Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Souhail ZAKI

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M. A. Sc. (Electrical Engineering)

GRADE - DEGREE

Department of Electrical Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Exploring Word and Sentence Similarity Corpus

C. Barrière

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

CO-DIRECTEUR DE LA THÈSE - THESIS CO-SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

J-P. Corriveau

N. Japkowicz

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Exploring word and sentence similarity in corpus

By
Souhail Zaki

A thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

M.A.Sc.
in

Electrical Engineering

Ottawa Carleton Institute for Electrical & Computer Engineering
School of Information and Technology Engineering (SITE)
University of Ottawa,
Ottawa, Ontario, Canada

December 2003

Your file Votre référence
ISBN: 0-494-01657-4
Our file Notre référence
ISBN: 0-494-01657-4

Abstract

This research addresses the problem of deriving semantic similarity between words of language using corpora and contextual distributions comparison methods. It aims to capture, in a comprehensive way, the similar behavior of words and henceforth properly estimates the semantic similarity between words of the language. The framework proposed for this purpose is incremental and iterative. The system combines the Edit distance and the incremental results as a way for accurate similarity measure. Moreover, a sentence similarity system is developed on top of the word similarity model. Naturally, the proposed model rests on observing the words behavior in large amount of natural text.

As for the strategy followed in this thesis, we first examine existing similarity measures, their hypotheses and show how these measures unfortunately fail to account for some linguistic features for estimating words similarity when they come under fine scrutiny. Furthermore, we present a model to enhance these measures to take into account linguistic characteristics.

Indeed, the suggested model takes large amount of raw data as input, extracts distributions of contexts and infers accordingly similarity between words using these distributions and Normalized Edit distance (NED).

Within this dissertation, we mainly report three implementations and evaluations of our proposal of similarity inference. The first implementation of the incremental inference model is applied to the cosine similarity measure. The enhanced similarity cosine is tested using a synonymy test. Secondly, the results of the first similarity measure improvement are used with the NED for estimating word similarity. Finally, a sentence similarity measure is proposed, implemented and tested using sentences form the same corpus.

Although, the original intent is similarity measurements for English words, the approach has eventually much wider applicability and extrapolations. First, this technique is language independent, the model and the implementation is independent from language, except for the corpora used for training. Similar corpora in different language will provide similarity measurement for words of that language with little or no alterations. Nonetheless, there are

few languages for which the order of the words is not rigid, and henceforth, the edit distance wouldn't be of great help to us, but these languages are but very few.

Acknowledgements

Many people are behind getting me to this stage and complete my research presented within this thesis. First, I strongly would like to emphasize that I was very fortunate to have Dr Caroline Barrière as a supervisor. I am unable to thank her enough for her support, encouragements and patience during my stay at Ottawa University and the trust she put in me. Very often, she would raise questions and observations that would open new venues of research for me.

Of course, my dear wife Khadija ElGharbi ought to have special appreciation for the tremendous support and endurance during my preparation. Her patience and encouragements while I have been away from campus is a key in getting me to this stage. This research is a gift for her as sign of my gratitude.

Special Thanks goes to my Mom and Dad for being supportive by reminding me all the times about how close I was to finish the endeavor I have started.

Immense gratitude goes to my mentor Sidi Hussein AbdelGawad for his fantastic wisdom that was a source of energy for me.

My dear friends Mouhcine Guennoun and Issam Jilani deserve extraordinary thanks for his unbelievable support and encouragements.

I would like also to thank member of my lab and colleagues Jérome tétreault and Mario Jarmasz for their fruitful discussions and help.

Of course, the list is very long; I would like to thank many other folks herein as well.

1 What is semantic similarity?

We know the semantic of a word or an expression has to do with its meaning. Several definitions of semantics have been offered throughout the linguistic and philosophical literatures, some of which are very close, whereas others are not. The very common definition, simply states the following:

"Semantics is the study of meanings"

Instead, other folks prefer to first relate linguistic expressions to meanings and then relate meanings to objects in the world as shown in this triangle [OGDE23] , commonly known as semiotic triangle:

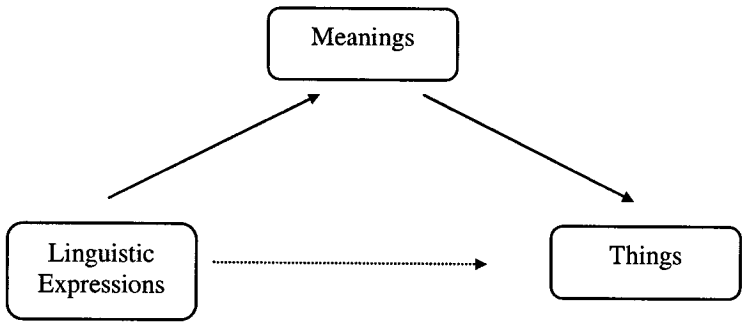

Figure 1 - Semiotic Triangle

"Linguistic expressions are related to meanings and meanings themselves are related to objects in the world. The task of semantics is to develop theories about the relations between expressions, meanings and the objects these meanings stand for" **[ALFA01]**.

The potential word of concern to us here is the word 'meaning'. It seems to be the core keyword. So how can 'meaning' be derived or measured through computational linguistics and Statistical Natural Language Processing? Is semantic similarity measurable?

The nature of similarity is a very debatable issue that is pointed out when looking for finding similar words or contexts. For some, similarity could be viewed as a distance measure. Conversely, not all people do share this Cartesian viewpoint. Indeed, for some people there is no real Euclidean space or distance and they propose other visions of similarity.

Miller & Charles **[MILL91]** state that semantic similarity is the degree of contextual interchangeability.

While measuring distances between concepts or words of language might seem a bit odd, it makes a great sense, when projected as follows:

The degree of relatedness of two concepts/words is correlated with the degree of relatedness of the contextual distributions of the concepts/words.

While automatically extracting the meaning of words is still a challenge, there are reasonable efforts for extracting semantic similarity between words of language. Therefore, most work on acquiring semantic properties of words has focused on semantic similarity acquisition; In other words, automatically acquiring a measure (metric) of how similar two words are. Indeed, countless and different approaches have been proposed by different schools working in the field of Natural Language Processing and Artificial Intelligence, with regard to the notion of the "semantic similarity", its identification and its characterization.

In this work, we will be looking at ways of deriving semantic similarity between words of language, enhancing them and improving on the best of them.

Natural Language Processing approaches to the study of meaning and words' similarity

The implicit assumption, upon which computational linguists base their investigations, is that the language has a regular structure. Therefore, every single word has its own features, semantic ones, syntactic ones. If one seeks to capture the syntactic-semantic features of a word, then one needs to count all paradigmatic and syntactic configurations in which the word partakes.

In our opinion, we believe that syntactic relations and semantic meanings are rather intertwined since the former ones are most of the time allowed by the latter ones.

Indeed, words of Language that appear in a sentence do have some regular interactions among each other. Those regularities and their discovery constitute the objective sought by computational linguists.

Waterman **[WATE94]**, in his work, performs a study about interactions between words and configurations in which these words occurs. More specifically his work seeks to characterize the behavior of a word with regard to other words, and to the various possible contexts in which it may occur.

This characterization is intended to enable judging which words are permitted in particular syntactic-semantic configurations and which are not, hence discovering selectional preferences/restrictions for words or even more for classes of words.

We do think that his work is endorsing our view of words' semantic in the sense that it clearly shows that set of words are naturally prone to appear in set of contexts. Even more, semantically close words tend to appear in the same contexts.

1.1 First school: paradigmatic relation[1], Aristotle way

1.1.1. Approach

According to Aristotle, the connotation of a complex concept may be defined in terms of concepts that are more primitive. Aristotle **[ARIS350]** once said: *'Man is a rational animal'*. Apparently, he defined the concept 'MAN' in terms of 'RATIONAL' and 'ANIMAL'. The type 'ANIMAL' is the genus or general type, and 'RATIONAL' is the differentia that distinguishes MAN from other types of ANIMAL.

Actually, 'RATIONAL' and 'ANIMAL' can themselves be defined in terms of still more primitive genera with appropriate differentiae until everything would be defined in terms of indivisible primitives.

Aristotle had divided the primitives, also called categories, into a set of eight different ultimate categories. These categories are Substance, Quantity, Relation, Time, Position, State, Activity, and Passivity. All the concepts can be, in the end, related to these ultimate primitives.

In his early works, Wittgenstein **[WITT72]** stated that compound propositions are made up of elementary propositions, which in turn are related to atomic facts about elementary objects in the world.

Taking this idea a step further, Masterman **[MAST61]** proposed semantic networks, a dictionary of 15,000 words defined in terms of 100 primitives.

1.1.2. Similarity: Path in a semantic network

Each concept basically, has two main parts: a genera and a differentia. Recursively, expanding each genus into finer genus and differentia will generate an Aristotelian hierarchy. Genera are the genus terms that function as the ancestors of a word sense. Differentiae denote the qualities that distinguish a particular sense from other senses of the same genus. Description logic is used to specify this differentia; it defines relations and selection constraints that are appropriate for each term, but that differentiate these terms from their immediate parents.

[1] A paradigmatic relation between lexemes is a pattern of association between two lexical units that share semantic information and have same semantic and syntactic roles. Traditionally, paradigmatic relations define semantic classes, syntactic categories, phonological natural classes, and distributional classes of all kinds.

A need to formalize these hierarchies for knowledge representation gave birth to semantic networks. In parallel, methods for knowledge inference, word disambiguation and other applications using these semantic networks have proliferated since.

The IS-A (or simply ISA) hierarchies have been a good example of the Aristotelian hierarchies. WordNet **[MILL91]** is a widely known semantic network project initiated by George Miller at Princeton University.

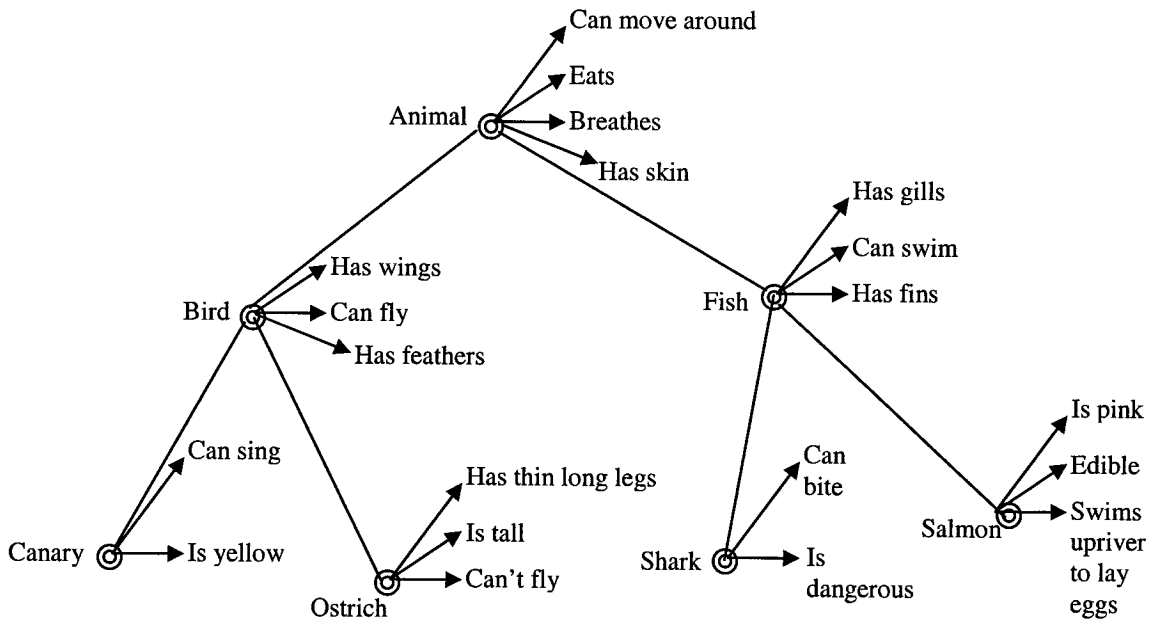

Figure 2-IS-A hierarchy as presented by Collins and Quillian (1969)

Evaluating semantic relatedness between two concepts (words) in an ISA hierarchy comes to measuring the distance between the nodes. The words will be semantically close if the distance (path) is short and remote otherwise.

Resnik **[RESN95]** defines a taxonomic similarity measure, which is based on the notion of information content. Under such view, semantic similarity between two words is represented by the entropy value of the most informative concept subsuming the two in a semantic hierarchy (WordNet). To give an illustration, all senses of the nouns '**clerk**' and '**salesperson**' in WordNet are connected to the first sense of the nouns '**employee**', '**worker**', '**person**' so as to indicate that '**clerk**'

and **'salesperson'** are a kind of employee which is a kind of worker which in turn is a kind of person. In this case, the semantic similarity between the words **'clerk'** and **'salesperson'** would correspond to the entropy value of employee which is the most informative (i.e. most specific) concept shared by the two words.

The distance is usually a mathematical equation of the links between nodes that form the path. Obviously, the problematic issue consists of rigorously assigning costs to links (edges). Most of the distances consider the costs as uniform across a hierarchy, which is a flawed hypothesis in our point of view. Richardson **[RICH97]** actually introduced a hierarchy measure where costs of the edges are not uniform; rather they are based the averaged probabilities of significance. Yet, even with such a more realistic cost assignment approach, there are still intrinsic issues with hierarchies:

- First, the kinships between elements of language (concepts, words) are subjective as it relies exclusively on human knowledge rather than objective linguistic data.
- Moreover, the costs of the edges, if any, will be likewise.
- Finally, this approach is not portable to other language unless hierarchies are built for every language, which is rather unjustifiably expensive and sometimes merely lacking.

1.2 Second school: Semes of features

1.2.1. Approach

Also known as, the componential analysis saw the light with Ferdinand de Saussure. He is considered as the founding father of the structural semantics. In a series of lectures at the University of Geneva **[SAUS16]**, he suggested that every language has a unique pattern and that the language consists of units. Moreover, the units can be identified only in terms of other units in the same language. To state the meaning of a semantically complex word we should try to give a paraphrase composed of words, which are simpler and easier to understand.

According to Saussure the phonemes constitute the minimal, atomic units that distinguish between words of language. He believed the structure of a linguistic system is one of combinations, contrasts, and oppositions, since the elements of language achieve meaning only in relationship.

Additionally, the componential approach dictates the meaning of a word is constructed out of elementary, smaller units of meaning, somewhat based on the analogy of the atomic structure of matter. These semantic atoms are widely known as semes, semantic features, semantic components, semantic markers, or simply semantic primes. These same atomic components are supposed to be indecomposable and universal.

According to Katz **[KATZ64]** the semantic structure of the semantic objects is encoded in their lexical entries.

For Katz, the semantic object 'Bachelor' is:

***Bachelor*:** **(Physical Object), (Living), (Human), (Adult), (Male), (Never Married)**

And for the woman:

Woman: **(Physical Object), (Living), (Human), (Female)**

They called the 'residue' of the word meaning after componential analysis 'the semantic distinguisher' and did not analyze that concept any further. Thus, one of the senses of the English word bachelor was represented by the set of componential features ('semantic markers' to Katz and Fodor) of (Human) (Adult) (Male) and the semantic distinguisher [Who has never married]. This meaning is, for Katz and Fodor, a combination of the meaning of man, derived fully componentially, and an unanalyzed residue. Katz and Fodor realized, of course, that such residue could be declared as another marker. However, this would have led to unconstrained proliferation of the markers, which would defeat the basic idea of componential analysis: describing many in terms of few.

It was shown, also by Katz and Fodor **[KATZ63]**, that the general lexicon could be represented using a limited number of semantic features.

On the other hand, the anthropologists Kroeber **[KROE52]**, Goodenough **[GOOD56]** and Lounsbury **[LOUN56]** suggested a set of semantic features (primitives) to describe terms of kinship in a variety of cultures. Using an appropriate combination of these features, one can compose the meaning of any kinship term. Thus, the meaning of 'father' is the combination of three feature-value pairs:

{GENERATION: -1; SEX: male; CLOSENESS-OF-RELATIONSHIP: direct}.

The approach could eventually be extended to cover every word of Language. This would effectively amount to the introduction of a 'meta-language' for describing word meaning, as relatively few features could be used in combinations to describe the hundreds of thousands of word meanings, presumably, in any language.

A more detailed and long-driven version of componential semantics is found in the work of Wierzbicka **[WIER72]**. Wierzbicka's view is that there exists a very restricted set of universal semantic atoms in terms of which conceivable meanings can be expressed. She has been working on

the universal set for few decades. Her inventory of primes is astonishingly small. She started out with eleven, but the list has now grown to sixty. A sample of the set comprises:

{I, YOU, SOMEONE, SOMETHING, THIS, HAPPEN, MOVE, KNOW, THINK, WANT, SAY, WHERE, WHEN, NOT, MAYBE, LIKE, KIND OF, PART OF, DO, BECAUSE, GOOD...etc}

Sowa in his book about **'Conceptual Structures'** notes that this is true even nowadays for dictionaries, with a slight difference:

"Modern dictionaries analyze thousands of words into more primitive ones, but they are not limited to a fixed set of categories. They also allow circular definitions: word A is defined in terms of B, which is directly or indirectly defined in terms of A." **[SOWA84]**

1.2.2. Similarity: Measuring of numbers of semes in common

Each concept is defined in terms of the smallest set of semantic primitives. The semantic primitives express what distinguishes one meaning from another. These same semantic primitives can also be used to form semantic classes, which correspond to groups of words that share semantic features and that can be exchanged in some contexts.

Each concept can be divided into ones that are more basic:

- **Man:**
  [+HUMAN]; [-FEMALE]; [+ADULT]
- **Boy:**
  [+HUMAN]; [-FEMALE]; [-ADULT]
- **Woman:**
  [+HUMAN]; [+FEMALE]; [+ADULT]
- **Girl:**
  [+HUMAN]; [+FEMALE]; [- ADULT]

Basically, the relation between two concepts is established by comparing their feature sets. Actually, this feature set can be viewed as a representational vector for the concept. Here again, all kind of conceivable vectorial measure would be able to measure the semantic similarity between words (tokens of language).

One famous similarity measure that is used is the following one:

$$\text{Jaccard coefficient} = \frac{|X \cap Y|}{|X \cup Y|}$$

Equation 1- The Jaccard Coefficient

Typically, the Jaccard coefficient is the number of shared attributes divided by the sum of unique attributes with regard to the two objects being compared.

On a parallel note, Tversky [TVER77] in his work about Features of similarity demonstrated that similarity between objects might be very well asymmetric. Consequently, he developed a more sophisticated formula for computing the similarity based on semantic primitives as:

$$sim(a,b) = \theta f(a \cap b) - \alpha f(a-b) - \beta f(b-a)$$

Equation 2 - Tversky similrity equation

Where:

- f(a∩b) counts the number of primitives that a and b have in common
- f(a – b) counts the number of primitives belonging to a but not to b
- f(b – a) counts the number of primitives belonging to b but not to a
- θ,α, and β are weightings indicating the relative importance of these three entities.

Nevertheless, the obvious problem raised by the componential analysis approach is the choice of primitives. All fellow researchers agree the primitives should be a small set of symbolic and atomic terms. Yet, there is no consensus on the set. Moreover, this approach is proven too restrictive and misleading sometimes. Indeed, there may be often several ways to decompose a concept/word that map to different contexts and situations.

1.3 Third school: context of usage and corpora

1.3.1. Approach

There are several tasks (text understanding, text summarization, sense disambiguation, information retrieval...etc) for which Statistical Natural Language Processing (SNLP) could make a big difference if we could automatically acquire meanings. However, how to represent the meaning in a way that can be understood by software agents is still a tough task. While some computational linguistics researchers argue that 'meaning' is something abstract, something like a 'shadow' of

words and sentences, that can't be computationally derived other folks insist it can be defined and characterized in numerous ways.

Adhering to the "Firthian" school: "You shall know a word by the company it keeps" **[FIRT57]**, one knows that he can get a thorough characterization of a word just by looking at the contexts in which the word appears.

In his late works, Wittgenstein states that "***to know the meaning of a word/sentence, we must look at how it is used and in what situation it is used***" **[WITT72]**

Likewise, Lyons in his book about linguistic semantics, offers similar statement where he says "There is no meaning in isolation; meanings arises from interconnection. " **[LYON95]**.

Actually, from a pragmatic point of view, we do know that the same word, changes behavior and meaning according to the contexts it appears in. Let us give an illustration of the phenomena:

I. "John will stay for a **month**"
II. "John will stay for the **month**".

Between the two sentences, there is a subtle change of meaning. The second sentence is introducing the fact that John is staying for the 'rest of the month'. It could be five days, it could be seven days or it could be more. We can effectively see the context influencing the meaning of the words and hence the meaning of a much larger unit which is the sentences of a text. Moreover, we can see that immediate context within a reasonable distance from target word have more influence and significance with respect to the meaning of that target word. Hence, in this research we are mostly interested in studying the semantic similarity between words of language based on the similarity of their immediate contexts distributions.

Let us give another example, this time of polysemous nature:

- Glass city is located at the **foot** of the Rocky Mountains.
- He scored a goal from a magical right **foot** shot.

The '**foot**' of the first example refers to the bottom part of the Rocky Mountains in the first while in the second example it refers to the human membrane, which is also the bottom of the legs.

So after tackling the aspect context of usage of words, let us examine the statistical methods for linguistics, which in the span of the last 2 decades or so have gone from virtually being unheard of to being a fundamental tool for linguists. Indeed, in recent years, statistical based semantic approaches have been widely used in an attempt to derive meaning form the large amount of available corpora. Linguists have been persevering looking for ways to leverage these huge amounts of electronically available texts from real daily language and from the internet. The potential of the huge amount data helping information retrieval and natural language exploration was so obvious that NLP folks kept delving into that direction. Of course, this theory could not have flourished before the advent of powerful computing machines and it took ampler acceptance and support with the proliferation of the internet and the availability of phenomenal amounts of texts. Certainly, the huge corpora available nowadays allow us to observe the language's words' usage in their natural environment. Thus, we get a more accurate idea about the constraints governing the words' usage and the configurations the words partake in. That said, any attempt to manually create all the examples of use would require phenomenally enormous and expensive effort if possible at all.

1.3.2. Similarity: Contexts distributions comparison

When carefully examined, the corpora-based approach seems more of an empirical approach while previous approaches seem more or less rationalist and theoretical ones. We frankly think that this particular point may be the source of attraction to corpora-based NLP explorations. Indeed, being able to start simply from empirical data and extract linguistic properties that are supposed to be in the mind of linguists is rather exquisite. These explorations started to flourish and prosper particularly after the mid 80, when more computing power and electronic literature started to materialize. One of the early uses of statistical studies was to study the word co-occurrence in large corpora **[SINC66]**, **[CHUR90]**. More attempts have been applied in order to model other linguistic features such as semantic classes, word sense disambiguation and more.

This thesis research is mostly concerned with approaches to semantic similarity problems in Natural Language Processing. Indeed, we are attracted in approaches that visualize words of language as having two intertwined features: Meaning and the distribution of contexts in which they occur.

Therefore, we second that words of language can be related to each other not only using their meanings but sometimes identifying the meaning relatedness requires looking at the words they co-

occur with. The semantic content of a target word '*w*' can, to a large extent, be characterized in terms of how that word goes together (co-occurs) with other words in everyday usage of the language. This everyday usage can be gathered through a large corpus. The semantic similarity between two words w_1 and w_2 is computed because of the extent to which their distributions of contexts of use overlap.

Although a large number of semantic measures and approaches have been offered throughout a huge research effort, the approaches differ in terms of:

i. How the notion of distribution is formally characterized (Euclidian distance, probabilistic framework, Information theory ...etc)
ii. The distance metric adopted to assess the proximity of two distributions.

Indeed many measures have been proposed and successfully used:

- In the work reported by Brown et al **[BROW92]**, the distance between two pairs of words is calculated using the averaged Mutual Information.
- In similar work, Grefenstette **[GREF92]** used the Jacquard coefficient described above **Error! Reference source not found.** to calculate similarity between distributions of contexts.
- Many other measures have been proposed as well.

1.4 Hybrid techniques

Basically, the idea is to combine the lexical hand coded resources with the empirical information from corpora.

The Linguistic String Project at the New York University performed one of the earliest attempts to categorize words according to semantic classes. Their work was based on the hypothesis that Harris relates meaning to distribution relative to other words as also, put:

"The meaning of entities and the meaning of grammatical relations among them is related to the restriction of combinations of these entities relative to other entities" **[HARR68]**.

Semantic classes were defined based on similar co-occurrence patterns of words within syntactic relations. However, the work relied on very small corpora and necessitated a lot of human effort. Indeed, Harris supervised a project named UNIVAC. It resulted in one of the first computer-based programs for performing syntactic analysis of English sentences. UNIVAC helped building a small dictionary with few ambiguities here and there.

Equally, Hearst **[HEAR92]** reported a method for "automatic" acquisition of the hyponymy lexical relation from unrestricted text. The method scans the text for instances of distinguished lexical-syntactic patterns that indicate the relation of interest.

An example of such patterns is:

... NP {, NP}* {,} **or other** NP...

An illustration of such pattern is:

"Bruises, wounds, broken bones or other injuries are common."

Then we would infer the interactions between entities as it follows:

Hyponym (bruise, injury)

Hyponym (wound, injury)

Hyponym (broken bone, injury)

One can occasionally see a semantic relatedness between entities tied with a hyponymic relationship to the same entity. This IS-A relation is a good indicator of meaning closeness and therefore can be thought of as an approach leading to semantic clustering and semantic discovery.

Grefenstette in **[GREF92]** pushes the idea even further by focusing his attention on the entities discovered by Hearst's technique. He proposes to deal with situations that may arise when one of the two entities (Hyponym/Hypernym) has several senses, especially, in situations where we want to add these entities to an existing network of lexical relations.

To illustrate more, let's say that we have discovered a new concept Hyponym (X, Y) but Y has multiple senses, then we should arbitrate to which sense it's more convenient to add the entity X. For instance, using WordNet, the task will be to determine which child subtree of Y, X shares contexts with. Grefenstette similarity hypothesis is:

"The context of a Hyponym X is significantly more similar to the contexts of the sense of Y that is its parent hypernym than to the contexts of the senses that are not its parent hypernym. Furthermore, theses contexts can be determined from the contexts that surrond the child subtrees of each sense of Y." **[GREF92]**.

Now, the question is to determine the appropriate context. Fundamentally, according to Grefenstette, the context is a bunch of features consisting of words in the corpus that are in syntactic

relation with the word (subject, object, adjective ...etc). Each context is grammatically analyzed yielding the stemmed form of the word and its syntactic role.

Then to compute similarity measure, Grefenstette introduced the Jacquard coefficient, which is the number of shared attributes divided by the sum of unique attributes with regard to the two objects being compared.

In other words, Grefenstette bases the similarity between words on their shared contexts. In fact, Grefenstette's technique is a hybrid method that combines language knowledge-intensive methods (extracting entities, parsing...etc) with statistical assessment.

Once again, while we think that such approach of combining the statistical information with the human classified information could lead to some acceptable results, we still think that human intervention is a major obstacle in our way to discover semantic relatedness in raw usage of the words.

1.4.1. Taxonomic techniques:

This may be a very intuitive semantic similarity that has been investigated in the late 80s. The first taxonomic relation that one can think of is the hypernymy/hyponymy semantic relation also known as the IS-A relation.

A natural way to evaluate semantic similarity in an IS-A network, like WordNet **[MILL91]**, is to estimate the distance between the nodes corresponding to the items (words) being compared. The shorter is the path from one node to another, the closest the two words are.

One situation that may arise, though, with the network representation is the existence of multiple paths. In theses circumstances one can adopt the shortest path approach Lee et al **[LEEH93]**.

Nevertheless, we think this view has an intrinsic shortcoming being that the edges between two nodes can't be precisely quantified with a uniform value.

Richardson **[RICH97]** working on the MindNet project at Microsoft, tried to take benefit of extensive linguistic parsing to assign different weights to different edges. While this approach provides reasonable ways of weighing the links between nodes, it still heavily relies on Knowledge base and human linguistic knowledge.

In the same way, in his work about Taxonomic similarity, Resnik **[RESN95]** proposes a richer semantic similarity in taxonomy, using the notion of information content. Letting C be the set of

concepts in taxonomy, the similarity of two concepts is the extent to which they share common information. In order to characterize the information of a concept, the taxonomy is augmented with a function:

P: C-------> [0, 1], such that for any c {belongs to} C,
p(c) is the probability of encountering an instance of concept c.

Following the line of reasoning of information theory [ROSS76], the information content of a concept c can be quantified as - log p(c).
It follows that this characterization provides a new way to evaluate semantic similarity.
"*The more information two concepts share in common, the more similar they are, and the information shared by two concepts is indicated by the information content of the concepts that subsume them in a taxonomy*" [RESN95].
Therefore:

SIM (c1, c2) = max [-log p(c)] | c∈ S(c1,c2)
With S(c1,c2) : Set of concepts that subsume both c1 and c2.

One question that remains is the estimation of the concept probabilities p(c).
Concept probabilities were computed simply as relative frequency, using a large corpus where each noun that occurred in the corpus was counted as an occurrence of each taxonomic.
Formally:

$$freq(c) = \sum_{n \in words(c)} count(c)$$

Equation 1 - Frequency equation

Where words(c) is the set of words subsumed by concept c. Concept probabilities were computed simply as relative frequency:

$$p(c) = \frac{freq(c)}{N}$$

Equation 2 - Probability formula

Where N is the total number of nouns observed.

Work by Barrière and Popowich [BARR00] clearly demonstrates that Resnik's approach is in fact limited. Actually the work goes on to show that all taxonomic based approaches are very limited because of the fact that the similarity measure must be expressed by an existing word in the language.

1.5 Who is right and who is wrong?

Naturally, there is no absolute answer to this though question. Usually, for one problem there could be more than one solution. Moreover, depending of the nature of the problem at hands (word sense disambiguation, lexicon generation, semantic similarity assessment, dictionary verification...etc) and depending on the precision required, many solutions can often lead to acceptable solutions to the same problem. A solid supporting proof is simply that most of the described approaches have been used extensively throughout the literature for different tasks and have been proven successful.

However, the study of relationship between contexts and words if approached by classical views can be extremely knowledge-intensive. One needs a great deal of linguists' expertise to be able to hand code the existing relations between words and contexts.

Similarly, we can effectively see that word meaning and semantic similarity between words of language are rather an empirical task. We also deem that the meaning conveyed by a word may be strongly impacted by its context. A context is not only a way to characterize the similarity between words but also a non-negligible part of the meaning of the word. In addition, the first two approaches assume that the meaning of a word has a context-independent value, which is in our opinion, definitely not accurate; In other words, to get an accurate idea about the constraints governing the words' usage and the configurations the words participate in, one needs to observe the behavior of words in a corpus. Conversely, a human effort to create taxonomies representing all the examples of use would require enormous and definitely expensive effort not to mention the limitedness of such endeavor.

Indeed, characterizations through statistical corpus studies are proven very precious and efficient for many reasons:

- Availability of data:

Nowadays, one can easily gather huge amounts of everyday language. Due to the over growth of internet usage corpora have been noticeably increasing in size, some of them reached 300M words. Yet, they are getting bigger and bigger.

- Knowledge-less approach:

One big asset of corpus-based techniques is the fact that one does not often need a lot of traditional knowledge input. Usually, knowledge-intensive techniques require expensive hand-labeled data (part of speech tagging, parsing...etc) or hand-coded rules (taxonomies, knowledge bases, real-world knowledge...etc) that you can get only at a great expense if possible at all.

- Objectivity:

Conceivably, when conclusions are based on hand-labeled data, we should expect to get biased by several choices about the input data and also sometimes by judgments about the results also. Dealing with plain empiric data, ensures at least that the yielded results are guaranteed to be bias-free.

- Automation and flexibility:

A similarly important reason is the fact that this technique relies only on raw data without any preprocessing or formatting, makes it automatic and language transparent; henceforth portable to other languages with little difficulties. We think that this make this kind of similarity measures very appealing from an information retrieval point of view as well as from linguistic common sense. Add to this, the fact that it can easily cope with the changes in the nature and size of the corpus.

Let's look at the following recapitulation chart:

Approaches / **Features**	Paradigmatic way similarity	Componential Analysis similarity	Corpora-derived similarity
Objectivity			X
Linguistic Knowledge Need	X	X	
Language transparency			X
Satisfactory results in general	X	X	X
Expensive data	X	X	
Empirical			X

Table 1- semantic similarity approaches comparison

1.6 Motivations and goals

To answer the question of why we do need to look at the word similarity, it is worth mentioning that word similarity could make a big difference for several applications. Below, we mention few of them for illustration purposes only.

1.6.1. Information Retrieval

Semantic similarity is used for information retrieval query expansion. When the user enters his/her request in terms of keywords, the system (search engine) can also suggest related searches based on similar keywords.

Another powerful side of using similarity is the K-nearest neighbors (KNN) classification. Let us say that we want to classify words according to the topic categories as they are used by newswire services.

The KNN performs in such a way to assign the new word to the most prevalent category among it's k nearest neighbors.

1.6.2. Dictionaries augmentation and verification

Another application of semantic similarity would be the verification of online dictionaries. Indeed, dictionaries have been built by human effort, so each dictionary has its own specific structure, words and gives specific explanations about the terms. In many situations, you can find definition of terms or synonyms that do not confirm to each other. Using such human-independent technique to assess the accuracy of dictionaries would be a very nice experience.

Furthermore, whenever new terms emerge in Language, proposed definitions are most of the time divergent. Finding semantically, syntactically similar words would make the defining process much clearer and easier.

That being said, there are plenty of other applications that would benefit from good word similarity measures. As an example, word sense disambiguation for automatic machine translation applications.

1.6.3. Goals

The central goals of this research [2] are to:

1. Thoroughly analyze the existing semantic measures and uncovering their shortcomings if any.
2. Propose and develop a sophisticated model of similarity to better capture semantic relatedness between words of language.
3. Eventually explore sentence similarity by using finer grained similarity measures adapted for linguistic flexibility.

Before developing a model that accounts for all linguistic features of words to extract semantic similarity, we deemed it was necessary to analyze the existing measures and observe their hypotheses and assumptions.

While this work examines existing similarity measures, it does not repudiate earlier work; rather it builds on top of it and fills some gaps left by fellow researchers. The other issue is the size of the context. Likewise, while this work picks a context of eight surrounding tokens as the best context for words, it does not fully address the issue. In fact, the 'best contexts' is by itself an excellent research venue that would be tackled in the future.

Another point we would like to emphasize is that we want to query raw corpora without any major linguistic preprocessing whenever possible. Indeed, we think that any preprocessing is an artificial layer that is forced unto the real language and may not reflect the usage of the words as they appear in everyday usage.

1.7 Contributions

The next two sections briefly outline the two major contributions of this thesis.

1.7.1. Revising a false hypothesis in corpus-based similarity measures

By now, we know that the notion of words' similarity ends up being the similarity of their contextual distributions often shown as vectorial representations or probabilistic representations. Stemming from our conviction that contexts should not be viewed with the binary vision: either equal or totally different, the first major contribution of this research is revisiting the hypothesis that

[2] Actually the original intent of the research was to develop a good semantic similarity measure in order to automatically extract 'causal patterns' from corpora; that is, using a handful of manually defined 'causal patterns' and bootstrapping using semantic measure, we would hopefully extract most if not all of the causal patterns present in natural language.

unfortunately serves as basis for existing similarity measures. Let us explicitly state the hypothesis upon which these measures rest:

"Two contexts are deemed related if they lexically match up; otherwise they are deemed as totally unrelated. When comparing two contexts, a word is totally similar to itself and totally dissimilar to all other words."

We believe that words of language can be semantically related without being the same lexicography. In fact, this is exactly what the whole semantic similarity endeavor is all about. In chapter 3, we will discuss this hypothesis accuracy and we will introduce some enhancements to cope with the hypothesis shortcomings.

1.7.2. A new similarity measure

Again, we think that taking into consideration tokens order is necessary while comparing contexts. Thus, the second major contribution is a similarity measure that combines state of art string dynamic comparison methods and the enhanced hypothesis above mentioned, to be able to extract most of the linguistic information out of contextual distributions. Chapter four presents in detail the rationale behind the new word similarity measure along with empirical implementations and discusses the results obtained throughout experiences.

The suggested similarity is a fine-grained similarity is the sense that it accounts for the ordering of the tokens in contexts. As previously mentioned defining, a better word similarity comes down to defining a better sequence similarity measure. Actually, the suggested similarity can be thought of as more of a sentential similarity measure.

1.7.3. A sentence similarity measure

While word similarity rests upon comparing contexts distributions, a better word similarity is a building block for sentence similarity. Thus, the other contribution is a sentence similarity measure that combines state of art word's similarity measure and the enhanced hypothesis above mentioned, to be able to estimate precisely sentence similarity. The new sentence measure along with empirical implementations is also presented in Chapter 4 in detail. A presentation of the results and discussion follows.

The suggested sentence similarity accounts for the ordering and insertions/deletions of the tokens in contexts.

1.8 Organization of the thesis

In the next chapter, we will examine most of the statistical semantic similarity and find out how these measures do compare contextual distributions and sequences of words. We will particularly explore measures rooted in vectorial and probabilistic philosophies. It also discusses the problems facing statistical explorations, which are challenging these similarity explorations. Chapter 3 exposes the binary hypothesis upon which the entire reviewed measures base their contextual comparison. Moreover, in the same chapter, we present a sound solution to the shortcomings of such hypothesis and we apply our solution to a particular vectorial measure, the cosine similarity measure. It also discusses evaluation of the measure and its performance. Chapter 4 presents a fine-grained word similarity using the Normalized Edit Distance. The new similarity measure is tested with a synonymy test for performance evaluation. The same chapter also delves into the fresh territory of sentence similarity. Chapter 5 presents a summary of the thesis and presents future work that is to be done in order to improve semantic similarity and avenues for further applications.

2 Related work

As presented earlier we have presented many approaches for deriving semantic similarity. While a lot of research in this direction relies on language rich methods for developing similarity measures, which make them applicable to a very limited number of languages, this work, will be focusing specifically on corpus-based, language-independent similarity approaches. We will examine approaches stemming from both vectorial vision of the world and probabilistic one.

With that objective in mind, the rest of this chapter is organized as follows. The next section discusses the topic of context of a word, how to determine the appropriate context and what linguistic information does the context enables us to extract. Section 2.2 describes the statistical techniques and their categorizations and particularly we will be interested in how these statistical techniques do perform contextual comparisons. Section 2.3 outlines some of the problems facing statistical approaches and how they have been approached. Section 2.4 delves into the methods for evaluating existing and probably future similarity measures. Finally, section 2.5 presents issues with existing similarity measures and how we intend to cope with these issues.

2.1 Context definition

A potential key word of concern to us comes into view: 'context'. It is supposed to be a configuration of tokens in which some target word appears. The common representation of a context is that of a vector of words co-occurring in the neighborhood of a target word.

The binary view of this vector catches the event (presence/absence) only. On the other hand the probabilistic view of the same vector is weighted by the frequency of vector/word co-occurring.

A hot debate is being waged about the 'Optimal' size the context should have:

- On side, very small contexts (1 or 2 closest words) neglect important interactions that may appear far away from the immediate neighborhood of the target word.
- On the other side, very long ones (The whole paragraph, One hundreds words...etc) would include all kind of noisy information in such a way that we will loose all kind of intrinsic information about the word in question.

In any case, a medium context from 3 to 6 tokens seems to be optimal as that's what is being used by many NLP fellow researchers. As stated previously, questioning such choice is outside the scope of this work. It will eventually, be dealt with in some future work.

2.1.1. Words configurations:

Let the following frame be a context:

The tall ... walked quickly to the bus.

There are few words or set of words, which might fill the gap.
For example:

"The tall **man** walked quickly to the bus. "

May be tens or hundreds even of words can go there; still it's a very small set of words with regard to the whole vocabulary.
We can see that the context influences the choice of the words and that not only one word but a class of words (woman, guy, policeman ...etc) can suit the context.
Similarly, one can play slightly with the tokens of the context and preserve the same words or category of words allowed to fill the gap.
For example:

"A skinny man walked furiously toward the red car. "

As it's obvious, these changes have almost no effects on the category of words able to fill the gap. In the same way, we can find tens or hundreds of other frames that will convey the same information while allowing the same category [human] to fill the gap.

So we can see that group of words can be related to a group of frames.
"These combinatorial patterns tell us that the properties that constrain the combinations are not idiosyncratically associated with individual words, but group themselves in patterns across word classes" **[WATE94].**

2.1.2. Class of words and class of contexts:

A class-based model is a model that sees the words as classes, based on the classes of contexts they appear or allowed to appear into. Discovering the interactions between classes of words and classes of contexts is a precious aspiration.

One of the aims for pursuing such objective is to be able to predict empirically what traditional grammars and linguists mandate about words usage, their constraints and combinations. Indeed, such discoveries will definitely enable us to understand automatically the properties inherent to the words and their constraints, combinations in order to identify the limits on the combinations of word in the language. In other words, we would be able to establish which class of contexts is associated to which class of words that can fit into it, eventually each with a weight.

Equally, to a word class, we can think of associated weighted contexts in which they are allowed to appear. These contexts would ultimately form a class of contexts.

The advocates of such view inform us about the fact that a word can be member of many classes since many words have several meanings and can have several grammatical roles too.

2.1.3. Validity of Corpus-based similarity measure

One of the core questions one asks before we go into more details about corpus-based measures is the validity of such statistical semantic characterizations.

Well, we know that corpus-based similarity measures are based on contextual characterizations of words and the evaluation of the distances (similarities) between the resulting contexts. We, are many in computational linguistics community have been looking at the issue of how informative is a context in which a target word appears, about the meaning of that word. In other words, finding an answer for the following question:

"Is it possible to mathematically and reliably establish a "semantic similarity distance" between words of language, based on contextual information only?"

From our point of view, we think that:

"The meaning of words lies in the contexts in which the words are used and can be characterized by the distribution of these contexts"

It turns out, in actual fact, that the contexts in which words appear are very informative about the meaning of those words to the extent that it's possible to learn the meaning of a word from the linguistic surroundings, only. Here is an illustration:

Example:

- Couscous is everyone's favorite dish.
- Barakuda restaurant serves the best Couscous in town.
- Couscous is made out of semolina.
- Of course, Couscous can be served with meat or fish.

Based on these sentences only, one can infer that '**Couscous**' may be some kind of dish made from semolina. One can see effectively that as humans we can make a connection between the meaning of words and the context surrounding these words.

Most of the researchers effectively think so; In an investigation of the associative and semantic priming, Lund et al **[LUND95]** state: "**the semantic vectors that are extracted from corpus are cognitively plausible and incorporate higher level semantic information, that may in part, correspond to semantic category and semantic feature similarity**".

Miller and Charles **[MILL91]** express the relationship between contextual representation and lexical meaning in the following statement:

"**Two words are semantically similar to the extent that their contextual representations are similar**".

From our point view, we know, that if two words are semantically very similar, then they can be interchangeable (substitutable) in most of the linguistic contexts they appear in. Reversing that postulate, we can affirm that: "**if the distributions of linguistic contexts in which two target words appear are very similar, then one can straightly infer that the two words in question are semantically very close**".

Work in **[MACD97b]** thoroughly investigated the validity of the contextual representation and its validity. The researchers offer the proposition:

"**If meaning is closely tied to the linguistic context, then the similarity of meanings and similarity of contexts should co-vary**".

The same work argued that human similarity ratings were very solid. In particular, the experiences showed that they change so little in times. Therefore, the paper used the

standard cosine measure results against human similarity judgments. Indeed, it was found that the results from the two approaches perfectly co-vary: For pairs of words with high human similarity ratings, the cosine ratings were correspondingly high and vice-versa.

The paper's findings effectively verify that corpus-derived similarity corresponds to a certain degree to human semantic similarity. Nonetheless, the remarkable thing about corpus-derived similarity measures is that they require no pre-encoded linguistic knowledge and are language transparent as well.

Numerous works specify exactly whether they are looking for syntactic or semantic properties, but several others do not and put it simply as semantic similarity. We prefer to talk about semantic relatedness between words in general, as semantic, syntactic properties are intertwined, and it is hard to separate them. Indeed, for a word to be part of some configurations, not only must it have the right meaning but also the appropriate syntax. In fact, not all combinations of words make sense. Take for instance the sentence:

"Eat spring computer"

While all symbols in this sentence are individually acceptable, yet they do not form an adequate sentence from linguistic point of view. Conversely, even if some combinations are syntactically sound, semantically they obviously are not. Take for instance the legendary example of:

"Green colorless ideas sleep furiously".

Discussion

Having demonstrated in the previous chapter and in the previous section that corpora can brilliantly help us characterize words' usage and words' meaning, we surveyed the literature for statistical similarity methods. In this research, we will not investigate the issue of most optimal context but will use a context of 3 to 6 words for the reason that that is the empirical optimal context. We will be particularly interested in knowing how these similarity measures perform contextual comparison.

Is the comparison of contexts focusing only on finding lexically exact context? or are similar but not exact contexts taken into consideration as well? Is linguistic feature of synonymy accounted for while doing contextual comparisons?

The following sections present a detailed overview of the methods used by other fellow researchers in their goal to infer words' meaning and semantic similarity. We will present

the algorithms used in these similarity measures and examine how they perform sequence comparison.

2.2 *Vectorial measures:*

NLP researchers comparing semantic similarity between words of language found themselves comparing vectorial distributions of contexts for the two words. Indeed, measuring words similarity generally meant measuring vectorial contexts distributions similarity. Each word is represented by a vector in a multi-dimensional space. Hence, the analogy between Euclidian and geometrical vectorial distance measures was rather expected.

For example, one could represent a matrix with words as columns and contexts as rows and a_{ij} will be number of times word j occurs in context i.

At any rate, there are two ways of constructing the contextual distributions:

- Simply reckoning the presence/absence of a target word in all contexts (binary values 1 or 0).
- Richer representations aiming to represent words as vectors in word space. Entry b_{ij} reflects the number of times word w_j co-occurs with word w_i. From an informational point of view, it makes perfectly sense to compute the frequencies of co-occurrence. It is worth mentioning here that co-occurrence can be defined with respect to documents, paragraphs, sentences or smaller units.

Several measures have been proposed to capture the similarity between the vectors characterizing the words, using both above-mentioned ways.

2.2.1. Binary vector measures:

Certainly, the most intuitive way of capturing the co-occurrence events is representing vectors as binary values (0 and 1) and ignoring the counts. Using such simplistic assumption researchers have ended up with several measures:

- **Matching coefficient:**

Simple, it counts the number of dimensions on which both vectors are non zeros. Its big disadvantage is that it does not take into account the length of vectors and the total number of non-zero entries in each.

Assume that the set of non-zero words for the first word is X ={wordλ, wordδ, ...wordη} and for the second word is Y = {wordβ, wordα,..., wordθ} then

$$\text{Matching coefficient} = X \cap Y$$

Equation 3 - Matching coefficient

- **Dice Coefficient:**

This coefficient was introduced as a remedy for shortfalls of the previous one. While it resembles the previous it, however, differs in the sense that it normalizes it by dividing by the total number of non-zeros entries. The measure is multiplied by a factor of 2 in order to let it range from 0.0 to 1.0

$$\text{Dice coefficient} = \frac{2|X \cap Y|}{|X|+|Y|}$$

Equation 4 - Dice coefficient

- **Jaccard coefficient:**

Jaccard's coefficient expresses the degree of overlap between two sets X and Y as the proportion of the overlap from the whole: $|X \cap Y| \ / \ |X \cup Y|$.

While the coefficient ranges from 0.0 to 1.0 like the Dice coefficient, it however gives lower values to low-overlap classes.

$$\text{Jaccard coefficient} = \frac{|X \cap Y|}{|X \cup Y|}$$

Equation 5 - Jaccard coefficient

- **Overlap coefficient:**

The idea of the overlap coefficient is to determine the degree in which the sets X and Y overlap each other. It has a 1.0 value if every dimension with a non-zero value for the first vector is also non-zero for the second one vector or vice versa

$$\text{Overlap coefficient} = \frac{|X \cap Y|}{\min(|X|,|Y|)}$$

Equation 6 - Overlap Coefficient

- **Cosine coefficient:**

The cosine measure relates the overlap of the sets X and Y to their geometric average. While it turns out to be the same as the Dice coefficient when the two vectors have the same number of non-zero entries, the cosine is used because it penalizes less in cases where the number is very different.

$$\text{Cosine} = \frac{|X \cap Y|}{\sqrt{|X| \times |Y|}}$$

Equation 7 - Cosine coefficient

This is very useful in statistical NLP since we often compare words or objects with different amount of data, but we do not want to say that they are dissimilar because of that.

Discussion

For all the measure with binary-valued representations that we have seen so far, each context constitutes a totally different dimension regardless whether two contexts are semantically close or not. Indeed, contexts are blindly compared; they are either completely exact when they match up lexically or completely different otherwise. The same applies for tokens of contexts. How about measures with real-valued contextual distributions?

2.2.2. Real-valued Vector Space :

So far we have presented binary vectors which are not very informative since they offer only one bit of information (presence/absence) on each dimension.

Dealing with linguistic objects, most researchers agree it is more suitable to capture real-valued contextual distributions. A real-valued vector $\vec{x}$ of dimensionality n is a sequence of n real numbers, where x_i denotes the ith component (ith context) of $\vec{x}$.

It can be written as:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ .... \\ x_n \end{pmatrix}$$

The length (norm) of a vector is defined as:

$$| \vec{x} | = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Equation 8 - Vectorial Norm

- **Cosine distance**

The dot product is defined as:

$$\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i \times y_i$$

Equation 9 - Dot product

The cosine measures the cosine of the angle between the two vectors. It ranges from 1 for vectors pointing in the same direction over 0 for orthogonal vectors.

Therefore, for general vectors we can define cosine as:

$$\text{Cos}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|}$$

This definition highlights another interpretation of the cosine measure as the normalized correlation coefficient.

- **Euclidian distance:**

The Euclidean distance between two vectors measures how far they are in the vector space:

$$| \vec{x} - \vec{y} | = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Equation 10 - Euclidian distance

However, this Euclidean distance seems a bit non-sense because the vector elements are just frequencies, then no Euclidean distance can be defined over frequencies space. This distance is not used at all.

Discussion

Again, while the measures based on frequency information perform better for word comparison, they still suffer from the same inaccurate hypothesis; each context represents a

different dimension by itself regardless of other contexts that may be semantically and/or syntactically similar; Of course, the same applies for words making up contexts.

2.3 Probabilistic measures:

Another school for quantifying similarity opts for probabilistic distributional comparisons. In order to identify elements sharing similar behavior we need to look for distributions with common traits. That is to say, we are looking for similarity measure among distributions that will capture similar/dissimilar behavior between the compared distributions.

Again, various measures of similarity have been widely used throughout several experiments.

2.3.1. Mutual Information

Early in the 90's Brown et al **[BROW92]** presented the outcome of investigation of representing language by class-based n-gram models instead of simple word-based n-gram models.

In their attempt to build an n-gram language model, they used probabilities of sequences of word classes instead of sequences of individual words.

The idea was to cluster syntactically and semantically close words into clusters and estimating the probability of the next word using the probability of previous word classes.

As a consequent of this clustering, the number of parameters needed to evaluate would become significantly much smaller. Likewise, probability estimates for many unseen data would be obtained easily.

Rather than defining a direct similarity measure between words, their model uses an interesting global optimization of the partitioning.

Let w_i be a word and assume there is a partition of vocabulary V into C classes using a function π that maps w_i to a class.

We represent:

$$w_1^n = w_1 w_2 \cdots w_n$$

In the classical n-gram, we would have something like:

$$\Pr(w_1^n) = \Pr(w_1)\Pr(w_2|w_1)...\Pr(w_n|w_1^n)$$

With:

$$\Pr(w_n|w_1^{n-1}) = \frac{C(w_1^{n-1}w_n)}{\sum_w C(w_1^{n-1}w)}$$

w_1^{n-1} is called history whereas w_n is called prediction.

In class-based n-grams we will have in addition

$$\text{For } 1 \le k \le n \qquad \Pr(w_k|w_1^{k-1}) = \Pr(w_k|c_k)\Pr(c_k|c_1^{k-1}).$$

Clearly, this will remedy to the data sparseness problem in the sense that it will be possible to make predictions for histories not previously seen by assuming that they are similar to other histories that have been seen. Estimating the model parameters for every word and every context is a huge task. Let us assume that we have a vocabulary of size V words and V^{n+m} contexts (We assume a context is n words to the left and m words to the right). Thus, the number of parameters the word model requires would be $2V^{n+m+1}$. For V such that V is $0(10^5)$ and n=1, m=1 then the number of parameters to estimate would be of the order of $0(10^{15})$ which is clearly impractical.

Therefore, the recourse for abstraction techniques seems to be very useful. By opting for a class-based model, we are going to have ℓ clusters of words and δ clusters of contexts; we need only $2\ell\delta$ parameters, which is a considerable reduction.

The problem then was how to partition the vocabulary into semantically-syntactically-related clusters. The Mutual Information was the measure that led to such clustering by looking for the partition that maximizes the average MI.
The work proposed the following algorithm to locate the optimum partition:

1- Initially assign each word to a distinct class and compute the average MI between classes.
2- Merge that pair of classes for which loss in average MI is least.
3- Moving some words from one class to another may result in a larger average MI, thus cycle through the vocabulary, after having a set of classes, moving each word to the class for which resulting partition has the greatest average MI.
4- Stop when no word reassignment leads to potential partition with greater average MI.

Using this approach they have been able to divide a vocabulary of 260,741 words into a 1000 classes averaging 260 words per class. Here are some of the interesting classes they have found:

- ***{June March July April January December October November September August }***
- ***{people guys folks fellows CEOs chaps doubters commies unfortunates blokes }***
- ***{down backwards ashore sideways southward northward overboard aloft downwards adrift}***

However, a look at some randomly picked classes reveal:

- ***{systems magnetics loggers products coupler Econ databanks Centre inscriber correctors}***
- ***{industry producers makers fishery Arabia growers addiction medalist inhalation addict}***
- ***{brought moved opened picked caught tied gathered cleared hung lifted}***

2.3.2. Why class-based models?

In our point of view, it is needless to argue that studying language interactions starts by looking at behavior of classes of words as a whole and not studying each word in isolation. Waterman argues:

"By grouping constructions and words into classes which have common functional behavior , we can abstract away from any idiosyncratic behavior of individual words, and beg to see these more general patterns of behavior" **[WATE94]**.

Another asset of the class-model with regard to the data sparseness question is the fact that no matter how big the corpora is, it is quasi-impossible to cover or see all the combinations

allowed for any particular word or construction. Ultimately, class abstraction will enable us interpolate probabilities for unseen events.

2.3.3. Universal similarity

Dekang Lin in his work about a "universal" similarity measures **[DEKA98]**, stemming from information-theoretic roots, gives a more general definition of similarity, which can be used for, words as well as for other objects:

$$sim(A,B) = \frac{\log(P(common(A,B)))}{\log(P(description(A,B)))}$$

Where:

- Common (A, B) denotes a function that states commonalities between A and B.
- Description (A, B) denotes a function that describes what A and B are.

Using a corpus distribution and a parser, he applied his similarity definition to target words and was able to extract similarity information.

Actually, this work is very similar to work initiated by Greffenstate **[GREF92]**, except that the latter is geared toward a generalist similarity measure. Indeed, commonalities between two words can be shared contexts, shared syntactic information...etc.

2.3.4. Bi-distributional categorical model

The study of the bi-directional relationship between contexts and words can be conducted through statistical corpus investigation.

"Each word, due to its semantic properties and syntactic constraints, will be able to participate in only a limited set of contexts" **[WATE94]**

Using a very large corpus, we simply note which pairs of words and contexts appear, and keep a list along with frequencies.

To be able to see clearly similar facts about contexts and words, one needs a distributional model for contexts with regard to the words in the vocabulary and vice-versa.

"We can find all occurrences of the contexts and a corpus and identify the sets of words tha appear with them, in order to evaluate whether they do in fact project similar constraints on word use.

Similarly, we can find all the contexts in which a single word appears, to evaluate the features of the syntactic environment with that word." **[WATE94]**

The author defines the distributional coefficients:

$w_j(c_i): c \in C \rightarrow [0,1]$: A word w_j's preference for a context c_i.

$c_i(w_j): w \in W \rightarrow [0,1]$: A context c_i's preference for a word w_j.

Note:

One important thing to bear in mind along this research is that discovering similarities contains implicitly potential information with regard to discovering dissimilarities, i.e. words that do not share the same behavior and therefore can't appear in the same contexts. Waterman 1994, combines ideas from class-based models and mutual interactions, he opts for condensing these discovered similar objects into compact form by introducing classes of words and classes of contexts.

Again, such idea will make it possible to reduce the parameters needed to evaluate and help when coping with the data sparseness problem.

To form such categories Waterman uses two formal sets of functions: one to match lexical categories $\ell \in L$, with context categories $\gamma \in \Gamma$:

$\ell_i(\omega_j)$: Membership of ω_j in class $\ell_i \in L$

$\gamma_i(c_j)$: Membership of c_j in class $\gamma_i \in \Gamma$

The context of a word is n word to left and m word to the right:

$$c \overset{def}{=} [t_{i-n} - t_{i-1}, w_i, t_{i+1} - t_{i+m}]$$

The word association: the word's w_j association with the context c_i is

$$w_j(c_i) \overset{def}{=} p(c = c_i | w = w_j)$$

The distribution for w_j is $w_j(C) \overset{def}{=} [w_j(c_1), w_j(c_2), ..., w_j(c_N)]$

The context association: the context c_i's association with the word w_j is

$$c_i(w_j) \overset{def}{=} p(w = w_j | c = c_i)$$

The distribution over the vocabulary is $c_i(W) \overset{def}{=} [c_i(w_1), c_i(w_2), ..., c_i(w_V)]$

2.3.5. Kullback-Leibler divergence

Kullback-Leibler divergence is a standard information-theoretic measure of the dissimilarity between two probability mass functions **[COVE91]**.

Work by Pereira et al **[PERE93]** used information theoretic measures Kullback-Leibler distance. This kind of measures deals with how much loss we have between the observed and expected distributions.

As previously stated, each word is represented into the contexts space by a distribution showing his presence/absence in each context and the frequency of participation in case of presence. The Kullback-Leibler distance is defined as:

$$D(w1||w2) = D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Equation 11 - Kullback-Leibler distance

p, q being the distributions for the two words (w_1,w_2) being compared; x would range over the dimensions in each distribution.

This divergence measures how much information we loose when substituting the distribution of w_1 by the distribution of w_2. This KL has been proven efficient for several applications. However, it has two major widely known drawbacks:

- D is not symmetric at all. It is obvious that there is no reason for D (w_1| |w_2) = D (w_2| |w_1). This is definitely an odd situation because the likeness between two words or objects in general should be the same whether you started to measure it from w_1 or from w_2.
- D is not defined whenever $P_{cont}(w_2) = 0$ which is very likely to happen a lot because of two things :
  - ✓ Not every word is going to occur with every context, otherwise the Language would be haphazard and arbitrary and therefore the whole Natural Language Processing would be simply useless.

- ✓ Secondly, our estimates are most of the time based on Maximum Likelihood Estimate that are calculated using the corpus. Consequently, in no way it contains all the possible combinations. This second draw back is really a handicap to use such measure. But, as will be discussed in section about data Sparseness problem, to solve the second issue we can opt for a smoothing technique to re-estimate the zero-frequency events.

On the other hand, the D measure is a distance measure whereas we are looking for a similarity measure. A solution **[DAGA97]** would be to use:

$$SIM(w1,w2) = 10^{-\beta.D(w1,w2)}$$

Where β is a conveniently tuned parameter.

2.3.6. Total divergence to the Average

Coping with shortcomings of the KL divergence, Dagan et al [DAGA97] introduced a related measure, in fact based on, the total KL divergence by averaging the two distributions:

$$A(w1,w2) = D(w1\|(w1+w2)/2) + D(w2\|(w1+w2)/2)$$

Where $(w_1+w_2)/2$ stands for the distribution $(P_c(w1) + P_c(w2))/2$ for c ∈ ***Contexts***.

The major asset with this distance is the fact that it overcomes both handicaps of D.

- ➢ As for symmetry, it's de facto a direct result of the definition. Indeed, $A(w1,w2) = A(w2,w1)$ for every (w_1, w_2).
- ➢ Furthermore, with the notations :
  - $H(x) = -x\log x$
  - $P_1(c) = P_c(w1)$, $P_2(c) = P_c(w2)$
  - $C = \{c \in Contexts / P_1(c) > 0, P_2(c) > 0\}$

One can show with little difficulty that:

$$A(w1,w2) = \sum_{c \in C} \{H(p_1(c) + p_2(c)) - H(p_1(c)) - H(p_2(c))\} + 2\log 2$$

With the following inequality:

$$0 \leq A \leq 2\log 2$$

So we can see that in order to calculate A we need only c ∈ C.

Again, A is a distance; A derived similarity measure has been proposed [DAGA97] similar to the one in the previous section:

$$SIM(w1,w2) = 10^{-\beta A(w1,w2)}$$

2.3.7. Confusion probability:

Essen and Steinbiss **[ESSE92]** the semantic similarity between to words is rather the extent to which these two words are substitutable in contexts and configurations. Hence, they came up with a different measure called confusion probability to measure the substitutability between the two words:

$$P_{confusion}(w1|w2) = SIM(w1,w2) = \sum_{c} \frac{P_c(w1)P_c(w2)P(c)}{P(w1)}$$

Equation 12 - confusion probability

It estimates the probability that word w_2 can be substituted for word w_1.

Discussion

As we can effectively see, even for probabilistic measures, each context with different lexemes is considered a totally different dimension by itself. Indeed, regardless of linguistic synonymy or not, contexts are blindly compared; they are either exact or totally different. Similarly, words of language are compared the same fashion.

2.4 Hybrid techniques

Approaching the probabilistic distributions with a geometrical focus has given birth to similarity measures that came from the combination of both worlds.

2.4.1. L1 norm:

Also known as the "Manhatten" norm, L1 is defined as:

$$L_1(w1,w2) = \sum_c |P_c(w1) - P_c(w2)|$$

Working a little around this expression, it can be easily shown that L1 (w1, w2) depends only on common terms for which the mass function is not null:

$$L_1(w1,w2) = 2 - \sum_{c\in C} P_1(c) - \sum_{c\in C} P_2(c) + \sum_{c\in C} |P_1(c) - P_2(c)|$$

This time we may consider another way of deriving similarity:

$$SIM(w1,w2) = (2 - L_1(w1,w2))^{\beta}$$

2.5 Observations and generalities

2.5.1. Binary Hypothesis

In this sea of semantic similarity measures, we find that for all the measure each context represents a dimension in the contextual distribution space. Comparing contexts rests upon the binary hypothesis: contexts are either completely similar if they lexically match up or completely different otherwise. This hypothesis is even interpolated to the finer elements of contexts that are words making up contexts. This hypothesis reflects in fact negatively on the evaluation of semantic similarity between contexts and consequently similarity between words of language. This adds up to the problem of data sparseness that will be thoroughly discussed later on.

2.5.2. Data Sparseness Problem

While corpora are very informative with regard to statistical language learning, they contain samples of language usage, which enables us to study empirical data in order to come up with abstractions and linguistic features that govern the language.

However, the corpora suffer from a serious problem related to lack of immense information about the language.

For instance if we are seeking to reckon contexts of just 3-grams, no matter how big the corpora are, there is slim chances that many of the possible contexts will ever occur in whatever big corpus we have access to. Most of the tri-grams don't appear in the corpus even in the largest collections of data, and worse, most of the observed events occur only once.

Indeed, the available data to use is quite insufficient to contain all the language events. For example, let's assume a vocabulary of V = 10^5 words, the number of possible tri-grams (could be seen as context of 3 words) requires at least V^3 (10^{15}) combinations while the biggest corpora that we have are in the order of 100 millions words. Apparently, a big portion of the combinations is not present in the corpus at all and the most remaining part appears occasionally once or twice. With corpora of 10^8 words, at most we are going to see 10^8 configurations, which mean (10^{15} - 10^8) is not seen at all.

Anyhow, one thing worthy of mentioning hereafter is the fact that with a vocabulary of V, not all the combinations of 3 words are linguistically acceptable due to grammatical and semantic constraints that rule out a huge number of them.

[Computer sat eat]

is for example a combination that is grammatically unacceptable. Likewise,

[Boat drinks printers]

is grammatically sound but semantically ruled out.

Context frequency:

Another viewpoint about context probability could be drawn from the fact that the context is compound of words. Therefore, the expected occurrence of the context is in the order of magnitude of the product of the probabilities of the words (independence assumption)

$p(c) \approx \Pi p(w_i)$. Therefore, if one word's probability is very low then the probability of context $p(c)$ is guaranteed to be very low.

Comparing distributions:

Taking into account the incredibly low frequencies for words and contexts, comparing the distribution of the two words is not going to hold interesting results because of this low frequencies bottleneck. There will be few words whose contextual distributions are sufficiently similar.

[BROW92] reported that for 3-gram model, using a very large corpus of 3.7×10^8 words, only 7.5×10^7 of 1.8×10^{16} possible 3-grams appeared. Furthermore, of these that did appear 71% did only once.

So we can see that the data on which the comparison is done is meagerly informative. Indeed, how can we successfully characterize a word or a context that we have seen only once or not at all?

In a similar way, the collected statistics have quantization effects. Indeed, events that appear will be highly over-estimated while unseen events are going to be highly under-estimated. This misrepresentation of low-probability is a serious pitfall. Indeed, when one needs to predict events or extract regularities, the inferred conclusions have to be based on repeated observations not only on events that hardly happen.

Consequently, researchers have opted for ways to reduce this quantization effects by assigning a non-null probability to unseen events.

2.5.3. Smoothing techniques:

To be able to build probabilistic frameworks with the available information only, co-workers use smoothing techniques with the objective of re-estimating observed frequencies.

While studying co-occurrence events between word and contexts, the main problem resides mostly in the contexts' rarity. As a remedy, researchers thought of estimating the contexts'

events using the frequency distribution of individual words. In other words, unseen events are accorded a probability based on single word distributions.

If we consider the MLE (Maximum Likelihood Estimate) for the probability of a word-context pair (w, c), it is simply:

$$P_{MLE}(c|w) = \frac{count(w,c)}{count(w)}$$

$count(w,c)$ is the frequency of (w,c) in the corpus and $count(w)$ is the frequency of w. Nonetheless, for a huge number of unseen pairs, P_{MLE} is zero, which leads to misleading probabilities estimates.

An option will be to take the MLE as an initial estimate and adjust it so that the total estimated probability of pairs occurring in the corpus is less than one, leaving some probability mass for unseen pairs.

2.5.3.1 Non-linguistic techniques:

- ✓ Additive Smoothing:

Certainly, additive smoothing **[LIDS20]** is one of the simplest and intuitive methods for smoothing data. It pretends that a n-gram (context) co-occurs with the word in question δ more than it does:

$$P(w|c) = \frac{count(w,c) + \delta}{\sum_{w_i} count(w_i,c) + \delta|V|}$$

Where V is the vocabulary.

- ✓ Good-Turing Estimate:

Good-Turing estimate [GOOD53] computes the estimated frequencies for unseen events using frequencies of the events seen. The Good-Turing formula replaces the actual frequency $count(w,c)$ of a word-context pair with a discounted frequency $count^*(w,c)$.

The work states that a pair that occurs r times should be treated as if it had occurred r* times where:

$$r^* = (r+1)\frac{n_{r+1}}{n_r}$$

where n_r is the number of pairs that occurred exactly r times in the training data. We can see that the new estimate rely on the number of possible events and the number, which were observed. Besides, to get a probability from this count, a simple normalization is used: For a pair word-context that occurred r time,

$$P_{GT}(c|w) = \frac{r^*}{count(w)}$$

As a consequence, the estimated conditional probability of an unseen pair (w',c') is

$$P_{GT}(c'|w') = \frac{(n_1 / n_0)}{count(w')}$$

Consequently, independently from the context all the pair (w,c) will have the same probability which is sort of weakness of such technique.

- ✓ Jelinek and Mercer

The work of Jelinek and Mercer [JELI80] is a classical way of interpolation. It basically, tries to interpolate linearly the MLE estimates for pairs.

$$P_{JM}(c|w) = \lambda(w)P_{MLE}(c|w) + (1-\lambda(w))P_{MLE}(c)$$

Of course, the function $\lambda(w)$ ranges between 0 and 1. In fact, it reflects our confidence in the available information that we got, based on the occurrence of x. Indeed, if w occurs enough in the corpus than we have all reasons to think that MLE for pair (w,c) are reliable

and henceforth , we assign a high value to $\lambda(w)$.Thus, $P_{JM}(c|w)$ depends mostly on $P_{MLE}(c|w)$. Conversely, if w is relatively rare, then $P_{MLE}(c|w)$ have few chances to be accurate and therefore we decide to make $P_{MLE}(c)$ lead us to satisfactory information about $P_{MLE}(c|w)$.

✓ Katz back-off

Katz uses a discounting **approach** that consist of using reduced MLE for seen word pairs, while using the left over probability mass to model unseen pairs. Indeed, he kept Good-Turing estimate for seen pairs but proposed a new formula for unseen ones:

$$P_K(c|w) = \begin{cases} P_d(c|w) \quad if \quad count(w,c) > 0 \\ \alpha(w)P(c) \qquad othewise \end{cases}$$

P_d: Good-Turing discounted Estimate for seen pairs

$\alpha(w)$: A normalizing factor required ensuring that $\sum_c P_K(c|w) = 1$.

NOTE: In a work presented by Dagan et al, $P(c)$ is deemed unfair for estimating the probability of the pair, thus the redistributed similarity that will take into account both w and c.

✓ Church and Gale

(Church and Gale 1991) introduce enhanced Good-Turing to overcome the drawback of Good-Turing. They use in addition, distribution of single words in evaluating bi-grams (for us co –occurrence of a word and the context). An unseen event could be estimated assuming that its components are independent. The authors argue that the lower order probabilities is a good indicator, thus $p(c|w) \propto p(w).p(c)$

2.5.3.2 Similarity-based Techniques:

If w' is similar to w, then w' can give us information about unseen pairs in which w is involved. We can think of using a weighted average of the evidence dependently on the similarity to w.

We can estimate the probability of an unseen pair (w2, w1) as follows:

SIM (w1, w'1): Increasing similarity function

S (w1): Set of words most similar to w1.

$$P_{SIM}(w_2|w_1) = \sum_{w_1' \in S(w_1)} \frac{SIM(w_1|w_1')}{N(w_1)} P(w_2|w_1')$$

where $N(w) = \sum_{w_1' \in S(w_1)} SIM(w_1, w_1')$: normalizing factor.

One point that needs to be stopped at is the definition of $S(w_1)$: For some (Essen and Steinbiss)**[ESSE92]** and [Karov and Edelman 1996], it's simply the whole vocabulary or the set of possible words V1.

As a consequent of the hugeness of V, computations become unwieldy.

For others **[DAGA95b]**, using the k-nearest words or words for which $dissimilarity(w_1, w_1') < t$ where t is a threshold, was more suitable.

2.5.3.3 Mutual Information for sparse data:

Dagan et al **[DAGA95a]** presents a method based on mutual information for estimating the probability of unseen word pairs. They estimate the probability of unseen co-occurrence using the co-occurrences that contain similar words. Their assumption is that similar words co-occurrences have similar values of mutual information.

Two co-occurrence pairs (w_1, w_2) and (w_1', w_2') are similar if w_1' is similar to w_1 and w_2' is similar to w_2. One stronger case happens when the pair differs only in one word. By the way, it's natural to assume that Mutual Information of the two pairs will be similar since we assume that w_1 and w_1' have similar co-occurrence patterns, w_2 and w_2' as well.

- The left context similarity of w_1 and w_2 relative to w, termed $sim_L(w_1, w_2, w)$ is defined as: $$sim_L(w_1, w_2, w) = \frac{\min(I(w, w_1), I(w, w_2))}{\max(I(w, w_1), I(w, w_2))}$$

- Whereas the right context similarity of w_1 and w_2 relative to w, termed $sim_R(w, w_1, w_2)$ is defined as:

$$sim_R(w, w_1, w_2) = \frac{\min(I(w_1, w), I(w_2, w))}{\max(I(w_1, w), I(w_2, w))}$$

2.6 What is next?

By now, we have established the groundwork for this thesis research. Indeed, we pointed out the way existing measures handles contexts and words comparison when comparing contextual distributions. We have effectively showed that the presented similarity measures do use a binary hypothesis for comparing contexts and henceforth neglect a great deal of linguistic information by neglecting the similarity between contexts and words. We do believe this hypothesis is a serious obstacle in our way of deriving comprehensive semantic similarity estimations. Going forward, we will be mostly concerned with enhancing these semantic similarity measures for better results using the similarity features of Natural Language. We have also presented the challenge of data sparseness that is hampering similarity measures. We do think that it is a problem only when we do view every combination of words as a dimension by itself i.e. when using the binary hypothesis. We do believe that overcoming the binary hypothesis will yield far better semantic similarity measures along with overcoming a big portion of the data sparseness issue. Indeed, when a word appears in a certain configuration, then the semantically similar configurations are correctly taken into consideration as well when comparing contexts distributions. We will not henceforth propose a separate method for dealing with the data sparseness issue, as we think that the proposed solution largely address the issue.

We will confine our attention in the remaining of this thesis to tuning the contextual comparison, which will enhance similarity measures while providing at the same time a remedy for data sparseness.

In the next chapter, we will present an enhanced word similarity measure. Experiments for evaluating the proposed similarity measures are presented therein. Further contexts comparison enhancement based on Edit distance are presented later in Chapter 4. The proposed word and sentence similarity measures are tested using different tests.

3 Revising an inadequate hypothesis in corpus-based similarity measures

3.1 Introduction

Most of the statistical methods for quantifying semantic similarities that have been proposed are performing decently to a certain extent, if we exclude the fact that they are resting on an imperfect hypothesis. This chapter actually details such over-simplistic hypothesis and shows ways to improve on these existing similarity measures. Simultaneously, this chapter shows that such hypothesis is divergent from the goal researchers are after and how this hypothesis deepens the data sparseness problem. Our immediate goal will be to give an insight of this hypothesis, widely used by many similarity measures, as well as showing a way to overcome the shortcomings of such foundation. However, in this research work we do not investigate measures stemming from semantic networks or paradigmatic approaches.

Researchers, using any of the statistical methods, unfortunately make use, of an inconsistent hypothesis. Indeed, when comparing distributions of words' contexts, they assume that two contexts are considered either totally related if they lexically match up, or totally unrelated otherwise. Even worse, as comparing contexts usually turns up into comparing tokens that make up the contexts, the same inadequate hypothesis is used over again:

Words are either completely semantically related if they lexically match up or totally unrelated otherwise.

Obviously, the hypothesis is not very compatible with the motivation of establishing the relatedness between lexically different words. This hypothesis naturally influences the results of all the methods in the sense that those methods will overlook the linguistic relations between words that exist in natural language. Hence, the resulting semantic characterizations and estimations become de facto imperfect.

Our proposition is to overcome this uncompromising hypothesis by providing a meaningful way of comparing contexts. Furthermore, we show that the current state of semantic similarity methods is just an initial step of a more complete process that leads to a more comprehensive estimation of words' similarity.

The remaining of this chapter develops this idea and details our contribution in the following fashion: Section 2 shows the validity of contextual characterization looking at several corpus-based similarity measures. Section 3 looks at the hypothesis used by such measures and shows its shortcomings. The section also discusses the results yielded by those measures and proposes a way of improving them. Section 4 illustrates this improvement by presenting a modification to the cosine similarity measure. Section 5 shows some results of our proposed measure applied on a set of words from a corpus. Section 6 concludes.

3.2 Quantifying context and word similarity

In this section, we argue that all of the vectorial and most of the hybrid methods can be at the end projected to be one form or another of the cosine measure.

In corpus-based similarity, sharing contexts is a sign of similarity. Therefore, the pertinent issue at hand here that is comparison of words relatedness ends up being a comparison of contextual representations of words. Of course, each word is represented within a high-dimensional vector space where the dimensions are contexts. We already discussed that a context could simply contain one word of the vocabulary (one-word window) as it could contain a compound combination of words (multiple-words window).

Therefore, any inaccuracies or shortcomings in contexts similarity assessment will undoubtedly reflect negatively on the words similarity task and vice versa. We show furthermore, that all the similarity assessment methods inherently make use of an inappropriate hypothesis.

Actually, all of the vectorial measures mentioned in the previous chapter are based upon the intersection of vectors, making them all closely related with slim variations.

The cosine measures the angle between the vector representations of the two words. The two words will be considered very similar if the vectors point toward the same direction, i.e. the angle between them is zero.

Conclusively, all the vectorial measures can be related to the cosine with little difficulty. Still, the comparison of dimension, which represents contexts, is the binary vision of black and white.

Therefore, in Section 7, most of our experiments would be based on the cosine measure as being representative of all vectorial measures. The choice of cosine measure is also appreciated because of its ability to incorporate the frequency information, instead of the over simplistic presence/absence information. At the same time, work by Schutze in **[SCHU92]**, the simple counting of absence/presence of contexts has been argued to be very little informative and misrepresentive. Henceforth, the co-occurrence counts are generally considered instead to get a greater appreciation of words' patterns of usage in corpora that we absolutely share.

At any rate, the representation of words becomes:

$$\vec{w}_1 = \sum_{c \in C} f_{1c}.c \quad \vec{w}_2 = \sum_{c \in C} f_{2c}.c$$

such that:

C: The set of possible contexts.

f_{ic} : The number of times the word w_i appears with the context c.

Lastly, the cosine measure looks like:

$$SIM_{Cosine}(w_1, w_2) = \frac{\vec{w}_1 \cdot \vec{w}_2}{\sqrt{|\vec{w}_1| \times |\vec{w}_2|}}$$

3.3 The Hypothesis inconsistency

We have described the notion in corpus-based similarity of a word's representation by its co-occurrences with different contexts, and the notion of words' similarity correlating to the similarity of their representations often shown as vectorial representations or probabilistic representations. We have also presented multiple measures of such similarity and we now look at the shared false hypothesis made by all these measures.

In all the vectorial measures presented, two contexts are considered equivalent ($c_1 \equiv c_2$) if and only if they contain the same words in the same order. For methods, such as the Edit distance **[LEVE66]** that will be introduced later, the comparison does not require the strict same order however; two contexts are identical if and only if the words are the same. In the probability measures, when comparing two distributions, only the information for

matching contexts is taken into account, which supposes the same hypothesis again. Let us state more explicitly this hypothesis:

"Two contexts are deemed related if they lexically match up; otherwise they are deemed as totally unrelated. When comparing two contexts, a word is totally similar only to itself and totally dissimilar to all other words of the language".

Now, let us examine the effects of such inappropriate assumption:

- First, this hypothesis is quite constraining and hampering, in the sense that two words can be related without having the same lexeme. In fact, that is exactly what all NLP folks working with similarity measures are trying to establish: the similarity between two words that are lexically different. The implicit premise behind all these considerable efforts is the conviction that words do not behave totally idiosyncratically, and that there is a huge deal of behavior relatedness between words of language. Therefore, it is a clear contradiction for researchers working with this hypothesis to assume the words of language as being orthogonal to each other. For a more formal Euclidian view of such hypothesis, each word constitutes a dimension in itself that is absolutely orthogonal to all the other dimensions. Word w1 can be in the same direction as w2 only if they are equal. The initial hypothesis they start with is obviously incoherent with the target goal.
- The other problematic aspect of such assumption is that it worsens the data sparseness problem facing corpora-based studies, even further. As shown by many studies a lot of words and context do show up few times even in big corpora. If the comparison measures perform poorly then you have simply wasted the chance of getting benefit of the few occurrences of words and contexts. We really think that if we successfully seize the opportunity of word/context correctly then the data sparseness would not be big of issue. Unfortunately, the existing measures do compare contexts and words in unjust manner, which makes the sparseness of data a serious matter.

Let us restate more formally, what we call the Binary Hypothesis:

$$SIM(w_1, w_2) = \begin{cases} 1 \; if \; (w_1 \overset{lexically}{=} w_2 \\ 0 \quad otherwise \end{cases}$$

We can see that this goes totally against the conviction and motivation of all of us working in computational linguistics, that words bear some relatedness even if they are lexically different.

3.3.1. Illustration:

Let the following sentences be two simple sentences from the corpus.

The subroutine **yielded** unexpected results. (S1)
That procedure **returned** unforeseen outcomes. (S2)

Imagine that the target words to compare are the words in bold: yielded and yielded. And let us consider a context of 1 word to each side.

- $Context(yielded) = \{The, subroutine, un\exp ected, results\}$
- $Context(returned) = \{That, procedure, unforseen, outcomes\}$

Of course, $\{The, subroutine, un\exp ected, results\} \overset{lexically}{\neq} \{That, procedure, unforseen, outcomes\}$

Using any similarity measures from any paradigm will not help us get any benefit from these two sentences though they are unmistakably very similar and very informative. Indeed, using the Binary Hypothesis, current similarity measures would suggest that the words (**yielded, returned**) have little, not to say nothing, in common. In these circumstances, no matter how good the similarity measure is and no matter how comprehensive the context is, the outcome is still the same.

This is simply because of ignoring the linguistic properties of words synonymy and semantic relatedness. In general, this is the direct repercussion of ignoring linguistic property of words relatedness.

The only hope for the similarity measures resting on the Binary Hypothesis, assuming the sentence S1 is unchanged, is to find somewhere in a large corpora, sentences such as:

That subroutine **returned** unexpected outcomes.

Or even better:

The subroutine **returned** unexpected results.

We all know that one of major problems of corpora examination is that a great deal of words appears only few times if they appear at all. Extrapolating this issue to configurations, we know that specific configurations and contexts will appear rarely if any. Hence, relying on the situation above-mentioned to happen is very unlikely. At the same time, ignoring information from close but not exactly identical context is frankly disregarding the richness and intrinsic qualities of natural language, which we are after anyway. Had we had, while comparing S1 and S2, some source of knowledge to inform the similarity measures algorithms that (That, the), (subroutine, procedure), (unexpected, unforeseen) and (results, outcomes) are very similar then, the conclusions with regard to (returned, yielded) would have been away better.

3.4 Proposed solution

Two major questions rise up:

1- How can we get a better characterization of words that would lead to a better relatedness measure?

2- Why can such a false hypothesis still lead to some satisfactory results?

Addressing the first question, access to a priori knowledge about the relatedness of words would be a solution. The main question then is how do we get access to contexts and word similarity relatedness?

Source of linguistic knowledge

We agree that comparing contexts and tokens while doing distributional comparisons should encompass the relatedness between these tokens. We cannot afford to neglect such significant information, especially if there are ways to derive it. What are our choices?

We can either on an external source of linguistic knowledge to provide information about words relatedness such as similarity measures using WordNet and Roget knowledge bases. Using such methods would effectively help our cause, however it will compromise our objective of purist ways for deriving semantic similarity because of :

- The non-portability of such methods to other language.
- The huge cost of knowledge coding and updating.
- The subjectivity and biases embedded within any large knowledge base as result of human judgments at the time of coding.

We propose a different approach, a very simple one, yet based on a very deep-rooted scientific method, known as **bootstrapping**, that states the following:

- One is allowed to start with a slightly incomplete hypothesis to get some a priori similarity insights.
- These preliminary results must be injected in the process again to get a more comprehensive appreciation of the semantic similarity. This step is repeated until a stability point is found.

First time, the measures algorithms are generally unchanged except for storing the results of the run. Of course, subsequent runs algorithms have to be adapted slightly in order to take into account the information about words similarity from previous runs. Further runs are made until a stability point is reached using the modified measure.

What do we find in literature? The results presented in the literature are always from the first run based on the flawed hypothesis and are presented as final and complete. In fact, the results that one gets in the first run are just a preliminary attempt to get some insight but definitely not the final results. For sure, this a priori information is acceptable and sometimes can be enough for some kind of applications, but further runs, always using the previous results to find better results leads to a better characterization of words in the language. Indeed, after a single run, one finds that related words co-occur with the same related words. Obviously, a word is very much related to itself, so we can see that the hypothesis used is just a special case of the more general hypothesis "a word can be semantically related with another group words".

Answering the second question, it is therefore no surprise that we are still able to get some meaningful appreciation even if with such extreme hypothesis. In spite of that, the results obtained from the first run are far from being comprehensive, simply because they certainly capture a very small portion of the words' interactions, but still are far away from yielding complete characterizations.

3.5 The adapted cosine measure

To better demonstrate the idea, we will consider a simple similarity measure, the cosine measure, with a very simple context being one word to each side. As said previously, the cosine is a representative measure for the vectorial paradigm.

3.5.1. Choice of cosine measure

Indeed, the core of all of the vectorial measures is the intersection of the distributional vectors. All the vectorial measure can be related back to the cosine measure with little difficulty:

- $Matching(X,Y) = |X \cap Y| = \sqrt{|X \| y|}.\cos(X,Y) = \alpha_{xy}.\cos(X,Y)$
- $Dice(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|} = \frac{2\sqrt{|X \| Y|}}{|X|+|Y|}.\cos(X,Y) = \beta_{xy}.\cos(X,Y)$
- $Jaccard(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{\sqrt{|X \| Y|}}{|X \cup Y|}.\cos(X,Y) = \delta_{xy}.\cos(X,Y)$

- $Overlap(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)} = \frac{\sqrt{|X||Y|}}{\min(|X|,|Y|)}.\cos(X,Y) = \theta_{xy}.\cos(X,Y)$

All the measures, basically, are a slight version of the cosine measure. However, the cosine is the most precise measure for measuring the similarity between two vectors. The other benefit of cosine is that it produces normalized measures, which helps with consistent comparisons between words of language.

3.5.2. Adapting the cosine measure

When working with corpus, we always have the problem of sparse data. In fact, about 50% of the words in any corpus will occur only once or twice. Eventually, we will be interested only in words that show up in the corpus.

To obtain a co-occurrence Matrix, we used a moving window technique of size 2*N+1, and for each target word; basically taking N tokens to the right and N tokens to the left. The counting consists of advancing the window through the sentences of the corpus, extracting the target word and the contexts, incrementing the appropriate matrix cell count.

Having two words $\vec{w}_1 = \sum_{i \in C} f_{1i}.c_i$ and $\vec{w}_2 = \sum_{j \in C} f_{2j}.c_j$, we defined the cosine as the cosine angle between the two vectors, having in mind that each dimension (context) is not necessarily orthogonal to other dimensions (contexts) if they are not lexically equal. The figure below shows the algorithm and framework proposed and used:

- Start with the Binary Hypothesis.
- Use the simple cosine
- Store the information in a database.
- Until we reach a stability point do
    - Use the full cosine
    - Compute the new results based on the previous results.
    - Store the information into a database for future reuse.
- Output the results.

Figure 3 - Similarity bootstrapping algorithm

For each run k we define the similarity at the k iterations as:

$$SIM_{iter\_k}(w_1, w_2) = \cos(\vec{w}_1, \vec{w}_2) = \frac{\vec{w}_1 . \vec{w}_2}{\sqrt{|\vec{w}_1| \times |\vec{w}_2|}}$$

where we define the scalar product:

$$\vec{w}_1 . \vec{w}_2 = \left(\sum_{i \in C} f_{1i} . \vec{c}_i \right) . \left(\sum_{j \in C} f_{2j} . \vec{c}_j \right)$$

$$= \left(\sum_{i,j \in C} f_{1i} . f_{2j} . \vec{c}_i . \vec{c}_j \right)$$

Again, since the vectorial product of $\vec{c}_i$ by $\vec{c}_j$ measures the similarity between context $\vec{c}_i$ and context $\vec{c}_i$ we can write the pervious formulas as:

$$\vec{w}_1 . \vec{w}_2 = \sum_{i,j \in C} f_{1i} . f_{2j} . SIM(c_i, c_j)$$

If we use the Binary Hypothesis, we find cosine measure usually used by researchers:

$$\vec{w}_1 . \vec{w}_2 = \sum_{i \in C} f_{1i} . f_{2i}$$

At iteration k of the process of knowledge reuse we have the following:

$$SIM_{iter\_k}(w_1, w_2) = \frac{\sum_{i,j \in C} f_{1i} . f_{2j} . SIM(c_i, c_j)}{\sqrt{|\vec{w}_1| \times |\vec{w}_2|}}$$

Now the question is about the contexts' similarity. Provided that we have got some information from the previous run, it's supposed to help us better understand words interaction. In brief:

$$SIM(c_1, c_2) = SIM_{iter\_k-1}(c_1, c_2)$$

Therefore,
$$SIM_{iter\_k}(w_1, w_2) = \frac{\sum_{i,j \in C} f_{1i} \cdot f_{2j} \cdot SIM_{iter\_k-1}(c_i, c_j)}{\sqrt{|\vec{w}_1| \times |\vec{w}_2|}}$$

The first run will be noted as BinaryCosine and each iteration k is noted as EnlightenedCosine.

3.6 Evaluating and testing of Semantic similarity

Now that we have developed a "new approach" for measuring semantic relatedness between words of language, we normally need to proof our "claim" of meeting linguistic goals and objectives. That is, does a new proposed distance provide better comparing distributions of contexts more than other similarity measures do? Does using the enhanced hypothesis instead of the binary hypothesis do enable us to achieve better insight into words' relatedness? Is this "new measure" suitable to some applications in specific or to any application in general? Is there an objective way of validating the results of any of the measures being used by different researchers? Is there a unified framework for evaluating all kind of measures?

Such questions and several others is the kind of questions that arise and when evaluating performances of different measures. Actually, the assessment of similarity measures is a critic undertaking. Certainly, it is easy to claim that one's method is the best but it's not easy to prove it. Several tests have been used to evaluate the measures:

3.6.1. Word Sense Disambiguation Task

One way for performing evaluations, found in abundance in literature, uses a word sense disambiguation application to show the performance of the proposed semantic measures. In such applications, the semantic measure is presented with an ambiguous word in some context and is requested to use the context to point out the correct sense of the word. For example, a test would be to present the measure with the sentence:

"crossed the **bridge**".

The method is then responsible to determine whether "bridge" refers to:

- A structure that extends a route
- A liaison between concepts, idea…etc
- A financial bridge
- A dental bridge.
- A card game.
- The higher part of the nose
- …etc.

3.6.2. Human Intuition

For many folks, visual examination of the resulting similarities often reveals groups of words that are intuitively semantically related. In general, this represents a generally accepted way of evaluating the similarity measures.

3.6.3. Human judgment Task

The other common empirical methodology for similarity judgment has been ratings over an n-scale scheme. Subjects are asked to rate pairs of words resemblance over a scale. Afterwards, the ratings are averaged to end up with a meticulous measure of semantic relatedness between words.

A very interesting experiment conducted by Miller & Charles **[MILL91]** consisted of providing appropriate human subjects with data. For example in their experiment, 38 undergraduate students were given 30 pairs of nouns that were chosen to cover high, intermediate, and low levels of similarity, and asked to rate "similarity of meaning" for each pair on a scale from 0 (no similarity) to 4 (perfect synonymy).

The average rating for each pair represents a good estimate of how similar the two words are, according to human judgments.

Likewise, fellow researcher Scott Macdonald notes that people can very reliably judge the degree of semantic similarity. Moreover, he showed that similarity judgments are very consistent over time. Indeed, his work established that "the Pearson product-moment correlation coefficient between ratings of 30 word pairs, made by two different groups of people 25 years apart has been an astonishing 0.97 (p=0.01)" **[MACD97b]**.

3.6.4. Our decision

While this research work doesn't delve into detailed investigation of this very attractive subject, we opted for task disambiguation evaluation for validating the results of our work. Indeed, we think, that evaluation methods investigation form a nice research topic that we are envisioning to explore in future work. A new emerging way of evaluating similarity measures, named Synonymy Problems Solving **[JARM03]** is gaining momentum. In fact, we think this method is more appropriate since it really focuses on testing the core functionality of similarity measures, Moreover; it's an automated way of testing which gives us the luxury of language transparency. Furthermore, this method does not rely on expensive and eventually subjective human judgments. Henceforth, we opted for using this validation methodology for our experiments.

3.6.5. Synonymy problems evaluation

Basically, a test would involve presenting the measure with a standardized set of synonymy questions. Each question aims to identifying the synonym of a term where the correct synonym is one four possible choices. Such standardized tests have already been collected and are usually:

- TOEFL [LAND97]: Consist of 80 TOEFL questions provided by the Educational Testing Service via Thomas Landauer
- ESL [TATS98]: Consist of 50 ESL questions created by Donna Tatsuki for Japanese ESL students.
- RDWP [LEWI01]: Consist of 100 Reader's digest Word Problems questions gathered by Peter Turney.

We think that this method of testing bears more objectivity and removes any subjectivity that could occur with other methods of evaluation such as human judgment. The only weakness with the method is that sometimes the choices are not chosen very carefully. In some cases only one candidate is really close to the target word, the others are very far from the target word, which could make the choice easier sometimes.

3.7 Results

We now present some results of testing the adapted cosine measure on a set of words taken from 2 corpora, the first one on composting, and the second one is one child breastfeeding.

These corpora have been developed by I. Meyer and Caroline Barriere at the University of Ottawa in 1998, and are used with their permissions.

3.7.1. Materials and Procedure:

1.1.1.1. Corpora

The first thing to mention is the corpus pre-processing. We want to do as little processing as possible that requires no linguistic knowledge. Neither part-of-speech tagging nor lemmatization was performed. In addition, as previously said, we wanted to capture all the linguistic interactions in words usage. Thus, we transformed all characters to lower case. At the same time, we kept all the tokens, including functional words and punctuation tokens as well. Only the numeric tokens were transformed to a token NUMERIC.

For our experiments we have used 3 different corpora:

- **Composting corpus**: A specialized corpus that revolves around the theme of composting. It was provided by Dr Caroline Barriere and Dr I Meyer. The corpus consists of 110668 words.
- **Breast-feeding corpus**: A technical corpus discussing the breast-feeding notions. It has 765832 words in it.
- **A Gutenberg based corpus**: A general-purpose corpus, relying on novels from the Gutenberg project. The corpus consists of 1689965 words.

1.1.1.2. Context

The basic context for a targeted word is a window of eight tokens, the next four tokens to it on the left and right sides. We opt to enforce the order of the tokens while comparing context rather than using the bag of words approach. Indeed, the order of the token into our daily language usage is very important henceforth the context comparison should consider it. Our purposes are:

- Showing the drawbacks of binary hypothesis.
- Presenting the benefit of knowledge reuse.

Indeed, even with a relatively small context the algorithm is still able to capture a big portion of the linguistic features of words.

3.7.2. Results

Our results presented here is just a subset of the whole vocabulary that shows the improvements of the average similarity throughout the iterations.

In the first columns, results of the similarity estimation between words, based on the Binary Hypothesis. The same results serve as a building step for more comprehensive estimations.

The semantic similarity measures have to pick the closest synonym out of four candidate words.

3.7.3. The test

We queried the databases for the best candidate for each question:

- prominent : { conspicuous battered ancient mysterious }
- enormously: { tremendously appropriately uniquely decidedly}
- advent : { coming arrest financing stability}
- urgently : { desperately typically conceivably tentatively}
- advent : { coming arrest financing stability}
- concisely : { succinctly powerfully positively freely}
- showed : { demonstrated published repeated postponed}
- constantly : { continually instantly rapidly accidentally}
- make : {earn print trade borrow}
- debate : {argument war election competition}
- arranged p: {lanned explained studied discarded}
- distribute : {circulate commercialize research acknowledge}
- physician : {doctor chemist pharmacist nurse}
- essentially : {basically possibly eagerly ordinarily}
- keen : {sharp useful simple famous}
- slowly : {gradually rarely effectively continuously}
- tasks : {jobs customers materials shops
- generally : { broadly descriptively controversially accurately}
- resolved : { settled publicized forgotten examined}

- feasible : { possible permitted equitable evident}
- terminated : {ended posed postponed evaluated}
- sufficient : { enough recent physiological valuable}
- fashion : {manner ration fathom craze}

Here are the results. For each word/measure we present the chosen candidate wit a + sign if correct and – otherwise. If the similarity measure cannot decide between the candidates a – (dash) sign is displayed.

Measure / Question	BCos	EnCos1	EnCos2	EnCos3	EnCos4
Prominent	Ancient(-)	Ancient(-)	Ancient(-)	Ancient(-)	Conspicuous(+)
Enormously	Decidedly(-)	Decidedly(-)	Decidedly(-)	Decidedly(-)	Decidedly(-)
Urgently	Desperately(+)	Desperately(+)	Desperately(+)	Desperately(+)	Desperately(+)
Advent	Coming(+)	Coming(+)	Coming(+)	Coming(+)	Coming(+)
Concisely	Positively(-)	Freely(-)	Freely(-)	Freely(-)	Freely(-)
Showed	Repeated(-)	Published(-)	Published(-)	Published(-)	Demonstrated(+)
Constantly	Continually(+)	Continually(+)	Continually(+)	Continually(+)	Continually(+)
Make	Borrow(-)	Borrow(-)	Borrow(-)	Earn(+)	Earn(+)
Debate	Argument(+)	War(-)	War(-)	War(-)	War(-)
Arranged	Explained(-)	Studied(-)	Studied(-)	Planned(+)	Planned(+)
Distribute	Research(-)	Acknowledge(-)	Circulate(+)	Circulate(+)	Circulate(+)
Physician	Doctor(+)	Doctor(+)	Nurse(-)	Nurse(-)	Doctor(+)
Essentially	Eagerly(-)	Eagerly(-)	Eagerly(-)	Eagerly(-)	Eagerly(-)
Keen	Sharp(+)	Simple(-)	Simple(-)	Sharp(+)	Sharp(+)
slowly	Gradually(+)	Gradually(+)	Gradually(+)	Gradually(+)	Gradually(+)
Tasks	Materials(-)	Jobs(+)	Jobs(+)	Jobs(+)	Jobs(+)
Generally	Broadly(+)	Broadly(+)	Broadly(+)	Broadly(+)	Broadly(+)
resolved	Settled(+)	Settled(+)	Settled(+)	Settled(+)	Settled(+)
feasible	Permitted(-)	Possible(+)	Possible(+)	Possible(+)	Possible(+)
Terminated	-	-	-	-	-
Sufficient	Recent(-)	Enough(+)	Enough(+)	Enough(+)	Enough(+)
Fashion	Manner(+)	Manner(+)	Manner(+)	Manner(+)	Manner(+)
Accuracy	**45%**	**50%**	**50%**	**63%**	**77%**

Table 2 - Binary and Enlightened cosine experiment

We also show at the bottom the percentage of accuracy of each measure.

From the results it appears that considerable gains were obtained as an outcome of knowledge re-injection and after abandoning of the drastic binary hypothesis.

We stopped at the fourth iteration because we got no further improvements in fifth and sixth iterations; the results were almost the same and stopped improving. We cannot provide an explanation as to why the algorithm converges into the fourth iteration and whether this is intrinsic to the corpus used or to the selected similarity measure which is the geometrical cosine.

Equally, getting a precise evaluation of words similarity, henceforth, a precise characterization of word behavior is too dear if we take into consideration the famous data sparseness problem. The rarity of some events (appearance of a target word in a corpus) is

usually a serious impeder to a complete word characterization. Now, we can use the better characterization for words sufficiently seen in the corpus to get a more or less better characterization for non-sufficiently seen words.

Effect of the context:

We wanted to explore what would be the effect of the context on the algorithm itself and the ability to leverage the recursion if the context of a word is reduced to smallest context possible. Henceforth, we performed the same test using the same corpus and same test using a context of one token to the left and to the right.

Here are the accuracy results for the cosine measure and context of two tokens:

Measure	BlindCos	EnCos1	EnCos	EnCos	EnCos
Accuracy	40 %	36%	36%	40%	50 %

We do notice clearly that the size of the context used to construct the contextual distributions do limit the efficiency of the algorithm in terms of similarity evaluations. This is widely expected since the longer the contexts the more the algorithm leverages the previous results from previous runs and vice-versa, the shorter the contexts the more unlikely the algorithm benefits from previous results.

3.7.4. Discussion and Conclusions

We have to confess that the distributional similarity quantified and collected from corpora using the binary hypothesis has proven to provide insightful information about words relatedness. Nonetheless, this hypothesis is drastically neglecting a great deal of substantial contextual information.

Indeed, we noticed that around the forth/fifth iteration there is no more gains. The same results are returned by the iterations.

Equally, one thing to mention is that these context-based measures permitted us to discover syntactical similarities as well. Two synonyms that are syntactically similar will have a bigger score than if they are syntactically different.

Throughout the experiences, we observed many instances where the BinaryCosine picked the wrong candidate and only iterations later, we have seen the EnlightenedCosine finally pick the right candidate.

Having said that, the two major observations that did pop up when conducting the experiences were:

- The first run with enhanced hypothesis usually brings up a clear and significant leap towards better estimation of the semantic similarity between words from the BinaryCosine.
- The following runs do fine-tune the similarity until it stabilizes in usually 4th or 5th round.

As for the first observation, we think it's natural as we gain a lot of insight about word similarity when we use results from previous run.

Nonetheless, there were few misses as well even with the enlightened cosine. This is natural as the data we use is not perfect. In an ideal world, you would have a corpus where every word is sufficiently described with all the configurations that word is supposed to appear. In addition, we would have every token appearing exactly the same number of times. Unfortunately, this is not the case with our corpus and hence the measure failures that we have seen and that was very well expected.

To recapitulate, in the previous section, we presented an iterative, comprehensive algorithm that takes into account the relatedness between two words without necessarily

being lexically the same. Certainly, we have described a novel framework for comprehensive and intelligent contextual distributions comparison. The validity of such judicious approach has been demonstrated by its application to estimate similarity between words from corpus using the cosine measure. Yet, that similarity framework should still be applicable to most of the existing similarity measures and would yield far better results.

While re-injecting the similarity information back into the similarity measure iterations brought significant improvements to evaluation of semantic similarity, we thought, moving forward, it might be worthwhile if we can further fine tune the process by picking up for a fine grained way for comparing sentences of words. The direction of better sequences comparison and dynamic programming was the direction that seemed the most suitable and promising direction, particularly that not a lot of work has been done in this direction.

4 Chapter 4: A fine grained word and sentence similarity measure

In the previous chapter, we looked at how existing similarity measures do perform contextual comparisons. We also have showed how modifying the blind hypothesis led to better characterizations and similarities by considering words and contexts relatedness feature. In this chapter instead, we take the best results of the previous enhancement and we look at how we can further fine-tune the similarities between words. We intend to poke into sentence similarity as well.

We will present in details the rationale behind the fine-grained similarity measure, its implementation and some results as follows:

Section 4.1 presents constraints that govern the usage of words and contexts. Furthermore, the section debates how slight changes to phrases and sentences can put casual similarity measures in severe troubles. Section 4.2 discusses in details and digs into existing measures for comparing two sequences. Section 4.3 introduces the Edit distance and some of its enhanced forms, for instance the Normalized Edit Distance for word similarity. Section 4.4 takes semantic similarity a step further into a new territory: Sentence Similarity.

4.1 Contexts and constraints of the language

In our opinion, to be able to achieve the acquisition of lexical semantics from raw text, one needs to carefully understand the features of the language and incorporate these properties into the similarity measures. Neglecting semantic features when comparing words and contexts can be costly as demonstrated in the previous chapter.

We do believe there other linguistic features that need to be uncovered and taken into consideration for better word similarity estimations. In these circumstances, we will bound our interest to exploring properties of contexts, including but not limited to, the importance of order of words, the flexibility to add/remove adverbial, adjectival and other functional words. We will also see how we can effectively get most of the semantic properties out of the existing sentences.

Without a doubt, we humans, use the language as a device for communication and exchanging plenty of ideas. It is no surprise we can express the same or semantically close idea more or less using the same amount of lexemes, but different wordings.
On the other hand, to convey a specific idea there may be a minimum number of lexemes that has to be used. For instance, if someone wants to express the idea that he checks his e-mail, he would say:

- I check my e-mail.

or at least similar wordings. But from the other side, there is nothing that constrains us to stick to these exact words and that same sentence.
In fact, one can add, depending on the situation and the circumstances, a flavor to the sentence without changing the main idea.

For example, others may say:

- I, regularly, check my email.
- I check my email every now and then.
- I, constantly, check my email.
- I, routinely, check my personal email.
- I check, more than once a day, my email.
- I check my email very often.
- I do check my email.
- I check my email account.
- I, always check my email.
- I check my email, frequently.
- ...etc.

Without doubt, we can see that all these configurations allow the verb '**check**' without being literally bound to the same lexemes combination and/or order.
Moreover, we could have used a very different verb:

- I verify my email.

- I do verify my email.
- I, regularly, verify my email account.
- I, always, verify my email, to be frank.
- I do verify my email.
- ...etc.

The addition of adjectives, adverbs, and other linguistic components, yet frequently used, certainly does not affect the main idea expressed by these sentences.
Therefore, if we were to compute the distances between the following contexts:

- I ... my email.
- I, frequently ... my email.
- I, frequently ... my old email.

These distances should be very small, not to say insignificant. Unfortunately, most of the existing measures, if not all of them, do not allow for such insertions and deletions of words. Rather, comparison is performed either by ignoring the positions of word or by looking into the strict aligned positions:

- ✓ I check my personal email.
- ✓ I frequently check my email.

That is to say that these measures would:

- Compare '**I**' against '**frequently**'.
- Compare '**My**' against '**my**'.
- Compare '**Personal**' against '**email**'.

henceforth yielding unsuitable results.
Therefore, crucial to words' similarity is the notion of sentences comparison.
Previously, we argued that the resemblance of two words' contexts distributions is a good indicator that the words in question are semantically and syntactically very close. In the following section, we meticulously examine several possible ways of comparing sequences of words.

4.2 Measures

Most of the sequence comparison measures that we find in the literature, for various and for words' similarity applications, neglect these linguistic features. We surveyed most of them in order to see how they perform sequence comparisons:

4.2.1. Matching Coefficient:

The matching coefficient measure is one of the very first intuitive methods that were used for sequences comparison. It sees a context of tokens as a bunch (Set) of words, with no association whatsoever to each other. The similarity estimate is nothing but the intersection between the two contexts, i.e. the matching words in the two contexts.

Actually, the measure obviously has a number of drawbacks. First and foremost, from a linguistic point of view, by neglecting the order in which words appear this measure is allowing all kind of insertions/deletions, irrespective whether they are linguistically permissible or not. Such a measure does not preserve the linguistic structure/information of the context (sequence of words).

Furthermore, let A be a context of word *w1* and B be a context of word *w2*. Let us assume that A and B have the same tokens but in very different (At the extreme opposite) order, the Matching measure will consider the two as being identical. Obviously, this is wrong and unacceptable. This goes without mentioning cases where contexts contain several instances of a token; all of the instances would be viewed as one token.

Add to this, another shortcoming that is: most of times, use of Matching Coefficient method has been concentrating on open set of words and completely neglecting functional words which are in our view crucial in characterizing behavior of any word.

In fact, we believe that any linguistically meaningful measure must allow for flexibility in contexts comparison while preserving the properties of the language structure, especially the sequential aspect of the context.

Several enhancements have been applied to the Matching measure, resulting in many derived measures, such as Dice Measure, Jaccard Measure, and Overlap Measure. All but few suffer more or less from the same weaknesses of the Matching measure.

4.2.2. Aligned Pair-wise:

One big leap to cope with previous weaknesses has been introduced through Aligned pair-wise comparison, which takes into consideration the order of the tokens in the contexts. The measure only considers tokens at the same position in each context, i.e. if we are to view each context as an array of words, the comparison would be based on similarity of tokens at the same index. This restriction goes hand in hand with our linguistic awareness of the importance of the order of tokens.

Let A again be a context of w1, B be a context of w2 such that:

$$A = [a_1, ..., a_n] \quad B = [b_1, ..., b_n]$$

The aligned pair-wise similarity would be sum of similarities between two tokens at the same positions:

$$P_{Aligned} = \frac{1}{n} \sum_{i=1}^{n} sim(a_i, b_i) \qquad \text{Where} \qquad sim(a,b) = \begin{cases} 0 & \text{If a} \neq \text{b} \\ 1 & \text{If a} = \text{b} \end{cases}$$

The values of $P_{Aligned}$ are ranging from [0, 1]. For identical contexts, the measure yields a value of 1. For totally different contexts, value of 0 will be returned.

Now, let us have a look at the following example:

- The student regularly does check his email.
- That assiduous student regularly verifies his personal email.

A = [The, student, regularly, does].

B = [That, assiduous, student, regularly].

$$P_{Aligned}(A, B) = 0!$$

Obviously, the introduction of adjectives (Assiduous), which is casual in daily writing style, has caused the collapse of the alignment of the sequences, hence the collapse of the

similarity between the two sequences. However, our cognition and language knowledge inform us the two sequences are similar, quite similar in fact.

While it is true that the measure considers order of tokens, nevertheless it overlooks an intrinsic feature of language, which is making allowance for insertions and deletions of casual linguistic entities such as adjectives and adverbs. These entities add flavors to the sentences without much affecting the whole meaning of the sentence. In other words, while these linguistic entities do not affect much the constraints of the configurations in which target do appear.

Furthermore, we can see that while there is a little improvement over the Matching coefficient measure, the Aligned pair-wise measure fails to take into consideration common properties of daily language usage.

It is both worthy to mention and amazing to know that despite the failure of Aligned pair-wise measure to account for insertions, deletions and substitutions several researchers have been able to successfully use it in several work efforts.

4.2.3. Edit Distance

While the Aligned pair-wise measure accounts for the order of tokens when comparing contexts, the measure imposes some rigid constraints that are far away from reflecting real linguistic features and real words' usage.

To cope with such hard constraints, we need a flexible way of comparing contexts that will allocate room for insertions, deletions and substitutions. Henceforth, we inevitably need to move toward the 'Edit Distance' first introduced by Levenstein **[LEVE66]**. The edit distance is based on dynamic programming that takes into account that the order of the tokens while comparing tokens and does not require the strict sequential alignment. The Edit distance allows for insertions, deletions and substitutions of tokens.

Definition:

The Edit distance is the minimum number of insertions, deletions, and substitutions required to transform one sequence into the other.

Consider two sequences X and Y and two words (tokens), w1, and w2.

Insertion: $XY \rightarrow Xw_1Y$

Deletion: $Xw_1Y \rightarrow XY$

Substitution: $Xw_1Y \rightarrow Xw_2Y$

We assume that each operation whether insertion, deletion or substitution is assessed as a "penalty". The cost of the penalty is generally 1 but variations of the measure can associate different costs to different operations. It would make more sense to attribute a penalty cost less than one in cases of tokens substitution whenever the tokens are lexically, syntactically and semantically close to each other.

The edit distance is the short sequence of operations that transforms sequence X into sequence Y. This method can also be thought of as finding the shortest path in a graph.

Edit Distance Computation:

Let $X = (x_1,...,x_n)$ and $Y = (y_1,...,y_n)$ be two sequences with n word each. We'll use dynamic programming method to compute the edit distance. To achieve this objective, we add an initial "fake" token for the two sequences at position 0 such that $x_0 = y_0$.

If we put D (X, Y) is the distance between the sequences X and Y, then we can recursively compute the D (X, Y) using the following rule:

$$\begin{cases} d(x_0, y_0) = 0 \\ d(x_i, y_j) = \min \begin{cases} d(x_i, y_{j-1}) + d_{insert}(y_j) \\ d(x_{i-1}, y_j) + d_{insert}(x_j) \\ d(x_{i-1}, y_{j-1}) + d_{substitute}(x_i, y_j) \end{cases} \end{cases}$$

Figure 4- Edit Distance

Where $d_{insert}(u)$ is the cost for inserting or deleting u, and $d_{substitute}(u, v)$ is the cost of substituting u for v.

To illustrate more, let's give an example:

X: I usually do check my email

Y: I do check my email

		x0	x1	x2	x3	x4	x5	x6
		-	**I**	**usually**	**do**	**check**	**my**	**email**
Y0	-	0	1	2	3	4	5	6
Y1	**I**	1	0	1	2	3	4	5
Y2	**do**	2	1	2	1	2	3	4
Y3	**check**	3	2	3	2	1	2	3
Y4	**my**	4	3	4	3	2	1	2
Y5	**email**	5	4	5	4	3	2	1

We start from d(x0, y0) downward to D(x6, y5) using the minimum rule (Equation 4.1).

The best alignment that could be found between the two sequences (X, Y) using the edit distance is the following:

I	usually	do	Check	my	Email
I	-	DO	CHECK	MY	EMAIL

The Edit distance allowed deletion (or insertion, depends on the origin/destination sequence) of the word 'usually'.

Discussion:

We can see that all the above-mentioned measures try to capture some linguistic aspects when comparing two sequences. However, the Matching Coefficient and aligned pair-wise measure fail dramatically to grasp the similarity we expect to be in real data and real contexts usage. Indeed, the matching coefficient measure offers the quality of taking into

consideration comparisons between any pair of words (tokens) from the two sequences; however, this comes at the cost of a configurationally factor of Language: sequence order. In a similar manner, the aligned pair-wise measure, allows comparisons between words (tokens) that are strictly aligned while neglecting a great deal of language variations that involves eventual mis-alignments by means of introduction of benign adjectives, adverbs and other functional words.

Doubtless, the Edit Distance, on the other hand, brightly incorporates both the configurational information and tokens mis-alignments characteristic. In reality, like the aligned pair-wise measure, the Edit Distance will attribute a high value to structurally similar sequences. Furthermore, it will rationally penalize two sequences that present some mis-alignments or tokens substitutions, nevertheless without collapsing down at the first mis-alignment.

While it doesn't go all the way toward incorporating comparisons of all pairs of tokens, as the Matching coefficient measure does, which is not suitable in our opinion anyways, it still brings into play comparisons between all relevant pairs of tokens to decide the best minimal set of operations.

Evidently, the Edit distance is a greater enhancement in our commitment to discover all the linguistic syntactic and lexical interaction of natural language words.

This distance function and its dynamic programming solution were developed in the 60s within the fields of coding/cryptography theory **[LEVE66]** and speech recognition.

The primary objective in speech recognition is to compensate for different speeds of speaking and thus stretch or compress the string phonemes in order to find the best match. This is often called 'elastic matching'.

By analogy, we see the edit distance as the appropriate choice for sequences comparison for it allows different wordings and styling with respect to expressing ideas and thus stretches or compresses the sequences of tokens in order to find the best match.

4.2.4. Normalized Edit distance

There are numbers of well-known algorithms for computing edit distances **[WAGN74]** ; many of these algorithms find their usefulness in error checking, pattern recognition, speech recognition, so on and so forth. Another excellent review of such distance measures

and their applications have been presented by Hall & Dowling **[HALL80]** and in the book edited by Sankoff et al **[SANK83]**.

Nevertheless, the edit distances as defined in the Figure 4- Edit Distance and in these applications are not very suitable for the linguistic similarity since they lack some type of normalization that would appropriately rate the weight of the operations with respect to the size of the sequences, subjects to comparison.

Certainly, the edit distance is a practical measure for similarity of two strings, yet sometimes the lengths of the strings compared need to be taken into consideration. Indeed, it is unfair to consider two sequences of words of length 100 differing in 1 word like two sequences of words of length 2 differing in 1 word; In this case, obviously, the former sequences are almost identical while the latter are definitely not.

To our knowledge, the Edit distance usage for semantic similarity is very rare and the usage of its normalized version is simply new to the field. The benefit of Normalized Edit distance is that it considers the lengths. These particular benefits pushed us to use this similarity for our proposal.

Editing Path

An editing path P between two sequences X and Y, is a sequence of points or ordered pairs of integers $(i_k, j_k), 0 \leq k \leq n$ satisfying the following:

Example:

Let $X = (x_1,..., x_n)$ and $Y = (y_1,..., y_n)$ be two sequences with n word each.

Definition:

$$
\begin{cases}
(I) \begin{cases} 0 \le i_k \le n \\ 0 \le j_k \le n \\ (i_0, j_0) = (0,0) \\ (i_n, j_n) = (n,n) \end{cases} \\
(II) \begin{cases} 0 \le i_k - i_{k-1} \le 1 \\ 0 \le j_k - j_{k-1} \le 1 \end{cases} \\
(III)\ i_k - i_{k-1} + j_k - j_{k-1} \ge 1
\end{cases}
$$

Norm of an Editing Path:

Given an editing path $P = (x_0, y_0)...(x_n, y_n)$ let the length of P, L(P) as the number of elementary edit operations described by P. Let also, the weight of P, W(P) be:

$$W(P) = \sum_{k=1}^{n} \lambda(X_{i_{k-1}+1} \ldots i_k \rightarrow Y_{j_{k-1}+1} \ldots j_k)$$

The Normalized weight of a non-null path is

$$\hat{W}(P) = \frac{W(P)}{L(P)}$$

And the normalized edit distance between X and Y is defined as:

$$d(X,Y) = \min \{ \hat{W}(P) \mid P \text{ is an editing Path between X and Y} \}$$

Note: In work for hand-written character recognition, Marzal & vidal [MARZ93], showed how NED can't be obtained from simply "Post-normalized" Edit distance, where first the ordinary edit distance is computed and then divided it by the length of the corresponding editing path.

4.3 Our proposal: Combining previous linguistic knowledge & Dynamic programming

We have seen in the previous chapter, the full benefit of re-injecting the results into the algorithm of vectorial cosine.

Similarly, we think that performing the Normalized Edit Distance experience reusing the results of the last cosine should lead us to a fine grained, accurately tuned word similarity measure. This way we stay loyal to our approach of comparing sequences, stemming from our conviction that sequences shouldn't be viewed with the binary vision either equal or totally different, and the same principle still applies when it comes to the lower level of particles of contexts, words shouldn't be viewed with the same binary (black/white) vision as well. With this spirit, and taking into consideration encouraging results from the previous chapter for the cosine measure, we modified the edit distance to use the result of Enlightenedcosine4 iteration to estimate similarities between words of language.

4.3.1. Experimentation and discussion

We present some results of testing the raw normalized edit measure as well as the enhanced normalized edit distance on the same set of words used in previous chapter taken from the same corpus.

- **Materials and Procedure:**

Again, we feel the need to emphasis that keeping corpus as natural as possible by avoiding all kind of pre-processing will yield authentic similarities and better words' characterizations. Indeed, no linguistic knowledge was applied to the corpus. Neither part-of-speech tagging nor lemmatization was performed. On the other hand and as previously stated in previous chapters, we are aiming to capture all the linguistic interactions in words usage. Thus, we transformed all characters to lower case. At the same time, we kept all the corpus tokens, including functional words, and punctuation tokens. The only exceptions were the numeric tokens that were transformed to a same token: NUMERIC.

The basic context for a targeted word is composed of words next to it on the left and right sides. The context is limited in terms of 4 words each direction and hence we think it should incorporate enough information for our goal sought here:

- Using the state of art sequence comparison measure.
- Presenting the benefit of knowledge reuse.
- Combining different semantic similarity measures from different approaches.
- Showing the precision and efficiency of the Normalized Edit distance.

➢ **The Test:**

We used the same test used in the previous chapter with synonyms problems.

We present only the results from EnlightenedCosine4 and Enlightened Normalized Edit Distance. Again, for each word we present the chosen candidate wit a + sign if correct, – otherwise and a simple dash if the measure cannot decide.

Question \ Measure	EnCos4	EnNED
Prominent	Conspicuous(+)	Conspicuous(+)
Enormously	Decidedly(-)	Decidedly(-)
Urgently	Desperately(+)	Desperately(+)
Advent	Coming(+)	Coming(+)
Concisely	Freely(-)	Freely(-)
Showed	Demonstrated(+)	Demonstrated(+)
Constantly	Continually(+)	Continually(+)
Make	Earn(+)	Earn(+)
Debate	War(-)	War(-)
Arranged	Planned(+)	Planned(+)
Distribute	Circulate(+)	Circulate(+)
Physician	Doctor(+)	Doctor(+)
Essentially	Eagerly(-)	Eagerly(-)
Keen	Sharp(+)	Sharp(+)
slowly	Gradually(+)	Gradually(+)
Tasks	Jobs(+)	Jobs(+)
Generally	Broadly(+)	Broadly(+)
resolved	Settled(+)	Settled(+)
feasible	Possible(+)	Possible(+)
Terminated	-	Ended(+)
Sufficient	Enough(+)	Enough(+)
Fashion	Manner(+)	Manner(+)
Accuracy	**77%**	**81%**

Table 3 - Edit distance TOEFL experiment

4.3.2. Results and discussion

In this experiment, we tested how the Edit distance fared in the experiment of TOEFL questionnaire. From surveying the results in databases and from the test shown above, it appears that not very considerable gains were obtained from the EnlightenedCosine4 measure in chapter3. However, if we compare these results to the original BinaryCosine, the achievements are quite substantial.

There are two reasons why the NED similarity measure does not do as well as expected:

- The imperfectness of the results of the NED is due the rarity of some events (appearance of a target word in a corpus) which is usually a serious problem to a complete word characterization.
- The other reason is that the results of NED are supposed to take where the cosine left off and since the Cosine results are not perfectly correct always either, the NED gets affected as well.

Nonetheless, the overall results from Normalized Edit Distance are away much better than the existing similarity measures.

4.4 Moving to Sentence Similarity

In the previous section, we looked at how using Normalized Edit distance led to slightly better word similarity by considering words and contexts relatedness feature. We used a context of 4 tokens into each direction. In this section instead, we thought about using the whole sentences into which a word appear as a context for that word. Using a sentence as a context along with Normalized Edit Distance, we ended up with a sentence similarity. Actually, it is very interesting to note that so much work has been performed for word similarity whereas none has been done for sentence similarity. We look at how we can take the word similarity measures to the next level such that we can derive sentence similarity for sentences found in the corpus.

Sentence similarity is just not so easy to define; is it that the two sentences mean the same thing? Or is it that they use the same words? Alternatively, is it that the two sentences use the semantically close words?

We will use the NED measure for the sentences from the same corpus and present in details the results.

4.4.1. Materials and Procedure:

We used a corpus about composting[3] that is compound of sentences. We tried to avoid any kind of pre-processing in a step to keep the corpus as natural as possible. Indeed, no linguistic knowledge was applied to the corpus. Neither part-of-speech tagging nor lemmatization was performed. Again, we kept all the corpus tokens, including functional words, and punctuation tokens. The only exceptions were the numeric tokens that were transformed to a same token: NUMERIC.

4.4.2. The Test

- We used the composting corpus, where we had the program to pick randomly 10 (ten) sentences and computes the similarity between these ten sentences and all the other sentences present in the corpus.

We present the 10 sentences with the semantically closest other sentences found in the corpus.

Sentence 1	*Sentence 2*
find two wide - mouthed glass jars , like those mayonnaise comes in .	find four or five wide - mouthed glass jars .
add one cup of soil or sand to provide grit for digestive process of worms.	add two handfuls of soil to supply roughage for the worms .
protozoa protozoa are one - celled microscopic animals .	protozoa protozoa are the simplest form of animal organism .
compost variables are the factors affecting the speed of composting .	other variables affecting the speed of composting include temperature , surface area and volume .
mites : mites are the second most common invertebrate found in compost .	springtails : springtails are extremely numerous in compost .
they are especially important in the formation of humus .	compost variables are the factors affecting the speed of composting .
they are especially important in the formation of humus .	actinomycetes are especially important in the formation of humus .
they chew on decomposing plants , pollen , grains , and fungi .	prey include minute nematode worms , mites , larvae , and small earthworms .
NUMERIC image of several worm composting bins.	NUMERIC image of wood composting bin.
making your own compost completes a natural cycle in your garden .	aerate your compost once a week to control .

Table 4- Sentence similarity experiment

[3] We would like to thank I. Meyer and E. Marshman from the School of Translation and Interpretation, University of Ottawa, for gathering the compost corpus, and allowing us to use it. The corpus contains 5646 sentence.

4.4.3. Results and discussion

This work in sentence similarity is in an exploratory stage. Henceforth, the resulted presented below are all but complete. In this experiment, we tested how the Normalized Edit Distance performed in the experiment of sentence semantic similarity. Actually, it is very costly to compare all the sentences together because of the cost of dynamic programming. That is why we picked randomly 10 sentences and showed the closest sentences picked by the similarity measures. Overall, it appears that the results that were obtained are acceptable.

Still we expect that the resulted of NED similarity measure will be affected by the following few factors:

- The rarity of some events (appearance of configurations of word in a corpus) which is usually a serious issue with regard to sentence similarity as well.
- In the same corpus, for a sentence we might not find many semantically close sentences.
- The other reason is that the sentence NED is supposed to take where the word NED left off and since the word NED results are not perfectly correct always either, the sentence NED gets affected as well.

That said, the sentence NED could be extremely useful since it can give an estimation of similarity between sentences in a text, corpus...etc. The only issue with it is the cost of comparing each two words in terms of computation and time needed to run such experiments.

5 Conclusions and future work

5.1 Semantic similarity

Semantic similarity has grown into a major field of statistical natural language Processing. That is mainly because, semantic similarity is very important for many applications, lexicon generation, word sense disambiguation, machine translation and many more. Many of these applications still rely on manually constructed linguistic taxonomies and dictionaries build at an expensive cost. Under such circumstances, one of the main objectives of the research reported in this thesis is to push semantic similarity measures to a deeper level. First, we remarked that statistical approach was more appropriate and reusable path seeking semantic similarity.

Secondly, during our investigation of semantic similarity we uncovered a widely used hypothesis, which is improper in our point of view. Therefore, we reviewed the hypothesis and presented an alternative approach to empower existing similarity measures in order to really evaluate the semantic similarity between words of language. Indeed, though our experiments focused mainly on the major existing similarity measures, which included cosine, and Edit measure, it is obvious that the same approach could be interpolated to all other measures with little difficulty. Coming to a greater understanding of semantic relatedness was also a motivation of this endeavor.

In addition, in our review of existing measures, which carry out semantic relatedness, we pointed out linguistic features including, but not limited to, adjectives and adverbs insertions and deletions, blindly overseen by most measures. None of the approaches and systems captures the fine-grained behavioral relatedness between words of language. Finally, we decided, therefore, that a new model of similarity measure that would explicitly account for such linguistic features in a manner compatible with our linguistic knowledge and cognition would be suitable for both word and sentential similarity evaluation. To address this need, we integrated the state of art of sequences similarity measure with the enhanced hypothesis to give birth to a solid, respectable similarity measure.

5.2 Contributions of this thesis

5.2.1. Hypothesis of existing measures

Our investigation and thorough exploration of most of the existing semantic similarity measures has permitted us to achieve a comprehensive analysis the base hypothesis upon which rest every approach and every measure. Therefore, we have uncovered an "improper" hypothesis used by virtually all the similarity measures. This hypothesis assumes that two words are either similar if they are lexically equal or totally unrelated otherwise.

5.2.2. Revision and Enhancement of hypothesis

The "improper" hypothesis caused the existing measures to simply inappropriately quantify the semantic relatedness between tokens of language. Indeed, by overlooking semantic relatedness between words, the relatedness between contexts was indeed badly measured. Hence, the whole contextual distributions were improperly quantified. This hypothesis has been enhanced and reused to give birth to accurate estimations of the similarities between tokens of language. We incrementally injected our automatically acquired linguistic knowledge into some of the simplest methods and we got much better and much accurate characteristic of semantic similarity between words of language. Simultaneously, we evaluated this new enhanced hypothesis using the normalized cosine measure; the results obtained were conclusive.

5.2.3. Language model and a sentence similarity measure

First, we effectively argued that the language is a very flexible when it comes to the insertion/deletion of adverbs, adjectives and other linguistic entities. Additionally we uncovered, through our complete examination of existing semantic measures, that most of these measures unfortunately do not support such natural properties of language. Indeed, sequences were being compared most of the time without allowing insertions, deletions or substitutions. Consequently, we proposed a more flexible word and sentence similarity model where sequence comparisons take into account these linguistic features in addition to reusing enhanced results of words similarity. Under such circumstances, we elected the normalized Edit distance to accomplish semantic quantification for both words and sentences of language, while accounting for linguistic richness and flexibility. Even more, we have successfully incorporated this art of state sequence dynamic comparison method

with the enhanced and revised cosine measure to create a very accurate word similarity measure. Finally, we used the same model for evaluating sentence similarity from the same corpus.

5.2.4. Implementation

We implemented a well designed generic framework for semantic similarity measures. The system takes corpora as input and produces information in form of huge databases that can be queried about relatedness between words of language. The Similarity Measure Enhanced Framework (SMEF) has been designed in such a way to be easily reusable for all kind of measures with minimal changes to the code and absolutely no change to the whole framework. We extracted samples of words and we showed how the framework drastically enhanced the quantification of the semantic relatedness between these sample words. Yet, the database contains millions of records for pairs of words. The nice thing about SMEF is that is Object oriented and reusable. Every major component of the system is implemented as a separate self-contained java class. SMEF supplied an easy architecture to plug in different similarity measures. We implemented the Normalized Edit Distance Similarity as application of our theory of flexible similarity measures model.

5.3 Future directions

This research work has opened up many paths for future exploration and examination. Some of these reflect issues in the Statistical Natural language processing that have not been resolved. Other issues simply involve unmistakable extensions to the work or potential applications of the work.

5.3.1. Optimal context

Undoubtedly, most of the semantic information for tokens of language relies in their contexts. Extracting this information relies heavily on the accuracy of the characterizing contextual distributions. No matter how smart a semantic measure is, the accuracy of the information extracted only depends on the accuracy of the features used for describing the words. Therefore, we strongly think that finding out the optimal context for word is a very much-needed step toward a perfect extraction of semantic information from contextual distributions. The model that we present we experimented different values for contexts ranging from 3 tokens from each side to 6 tokens. We do think that the optimal values must

be in this range, yet we do not have any proof for this claim. Indeed, while our model uses the thought-to-be optimal contexts does not delve into this subject. The optimal context forms a good research spot.

5.3.2. Information Retrieval bootstrapping

We have mentioned in section 1.4 that some researchers use an hybrid approach to attempt at extracting concept related by specific semantic relations from corpus. These semantic relations, such as hypernymy, synonymy, function, causality, are defined by a set of patterns. For example, causality can be expressed by small phrase of few words like:

- Hence, therefore...etc.
- Because of, on account of...etc.
- Because, since, as...etc.
- That is why, the result was...etc.

Nevertheless, this set is limited, and it is very time consuming to search for new relation defining patterns in text. It would be interesting, as future work, to explore whether the semantic similarity measure defined here, could help find more patterns, similar to a predefined set of patterns.

Indeed, starting up with a handful of manually defined 'causal patterns' and bootstrapping using the proposed semantic measure we would reckon most if not all the causal patterns present in natural language.

5.3.3. Multiple Corpora

One of the options we would like to test is the effect of using different corpus at different iterations. We think that each different corpus brings different configurations and situations for words. Hence, in any new iteration we would take benefit from better similarity estimations and new words contextual information.

5.3.4. Evaluation

We are particularly very concerned and very interested in extending the current work by again a thorough examination and investigation of different evaluation methods for semantic similarity measures. We think, a priori, that some of the existing evaluation procedures might be very limited and sometimes non-conclusive. We would also like to improve these current evaluation methods. Finally, we are persuaded we can develop a

formal validation method that will objectively assess the accuracy of the measurements rendered by similarity measures.

5.4 Conclusion

The distributional descriptions of words of language, collected and measured from large bodies of text have been shown to contain extremely valuable semantic information. This work and embodied experiments within validated these concepts and proposed an enhancement to existing similarity features. Furthermore, this research goes one-step further in combining state of art of sequence similarity with enhanced hypothesis to bring into light a powerful similarity measure.

6 References

[ALFA01] Abu Nasr Mohammad Ibn al-Farakh Al farabi (c.872-950) Ihsa' al-'ulum (Enumeration of the Sciences).

[ALLE95] James Allen 1995, Natural Language Understanding. Benjamin/Cummings Publishing company, Inc., second edition.

[ARIS350] Aristotle, 350 B.C.E. *Categories.* Translated by George MacDonald Ross, 1975.

[BARR00] Barrière Caroline and Popowich F., 2000. *Expanding the type hierarchy with non-lexical concepts.* Proceedings of Canadian AI, Best Paper Award, p. 53-68 ,Springer Verlag.

[BROW92] Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class based n-gram models of natural language. Computational Linguistics, 18(4):467--479.

[CHEN96] Stanley F. Chen, Joshua Goodman 1996. An empirical Study of smoothing Techniques for language modeling. Proceedings of 34th Annual Meeting of ACL, June 1996.

[CHUR90] Kenneth Church, Patrick Hanks 1990. Words associations, Mutual information and Lexicography. Computational Linguistics, 16(1): 22-29.

[COVE91] Thomas M. Cover and Joy A. Thomas, 1991. Elements of Information theory. Wiley series in telecommunications. Wiley-Interscience, New York.

[DAGA95a] Ido Dagan, Fernando Pereira and Lillian Lee, 1994. *Similarity-based estimation for word co-occurrence probabilities.* In ACL 32, pp. 272-278.

[DAGA95b] Ido Dagan, Shaul Marcus, Shaul Markovitch 1995. *Contextual Word Similarity and Estimation from sparse data.* Compuer Speech and Language, 9 :123-152.

[DAGA97] Ido Dagan, Lillian Lee, Fernando Pereira, 1997. *Similarity-Based Methods for Word-sense Disambiguation.* In 35th Annual Meeting of the ACL, pages 56-63, Madrid, Spain, July. Association for Computational Linguistics.

[DEKA98] Dekang Lin, 1998. *An Information-Theoretic Definition of Similarity.* Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July, 1998.

[ESSE92] Ute Essen and Volker Steinbiss, 1992. *Co-occurrence smoothing for stochastic language modeling*. ICASSP 92, volume 1, pages 161-164.

[FIRT57] Firth, J. R, 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

[GOOD53] I. J. Good, 1953. *The population frequencies of species and the estimation of population parameters*. Biometrika, 40:16-264.

[GOOD56] Goodenough, Ward, 1956. *Componential analysis and the study of meaning*. Language 32. 195-216.

[GREF92] Gregory Grefenstette, Marti A. Hearst, 1992. *A Method for refining automatically-discovered lexical relations: Combining weak techniques for Stronger Results*. Papers from the 1992 Workshop Technical Report W-92-01, AAAI Press, Menlo Park, CA, 1992.

[HALL80] P. A. V. Hall and G. R. Dowling, 1980. *Approximate string matching*. Computing Surveys, 12(4), 381-402.

[HARR68] Zellig Harris, 1968. *Distibutional hypothesis*.

[HEAR92] Marti Hearst, 1992. *Automatic aquisition for Hyponyms from large text corpora*. Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, July 1992.

[JARM03] Mario Jarmasz, 2003. Roget's thesaurus as a lexical resource for Natural Language Processing, Master thesis, University of Ottawa.

[KATZ63] Katz, J. and Fodor J., 1963. *The structure of a semantic theory*. Language 39: 170-210.

[KATZ64] Katz, Jerrold J, 1964. Semantic *Theory and the Meaning of 'Good'*. The Journal of Philosophy 61, no. 23: 739-766.

[KROE52] Kroeber, A.L. & Kluckhohn, C. (1952): Culture - A Critical Review of Concepts and Definitions, Peabody Museum Papers 47,1 Cambridge, Mass - Harvard University Press.

[LEEH93] Joon Ho Lee, Myoung Ho Kim, and Yoon Joon Lee. *Informaion retrieval based on conceptual distance in IS-A hierarchies*. Journal of Documentation, 49(2):188-2007, June 1993.

[LEEL97], Lillian Jane Lee 1997. *Similarity-Based Approach to Natural Language Processing*. PHD thesis, harvard University, Cambridge Massachusetts.

[LEVE66] A. Levenshtein, 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Phys. Dokl., 10:707-710, 1966.

[LINF96] Francis Y. Lin, 1996. The Syntax, Semantics, and Inference Mechanism of Natural Language. Knowledge Representation Systems based on natural language, AAAI-96 Fall Symposium series, MIT.

[LIDS20] G.J. Lidstone, 1920. *Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities*. Transactions of the faculty of Actuaries, 8:182-192.

[LOUN56] Lounsbury, Floyd G. 1956. A semantic analysis of Pawnee kinship usage. Language 32. 158-194.

[LINY98] Francis Y. Lin 1996. *The Syntax, Semantics, and Inference Mechanism of Natural Language*. Knowledge Representation Systems based on natural language, AAAI-96 Fall Symposium series, MIT.

[LUND95] Kevin Lund, Curt Burgess and Ruth A. Atchley, 1996. *Semantic and associative priming in High-dimensional semantic space.* In proceedings of the 17th Annual Conference of the Cognitive Science Society. Pittsburgh, PA, 660-665.

[LYON95] John Lyons, 1995. *Linguistic Semantics: An Introduction.* Cambridge University Press.

[MACD97a] Scott McDonald, 1997. *Exploring the validity of corpus-derived measures of semantic similarity.* Paper presented at the *9th Annual CCS/HCRC Postgraduate Conference*, University of Edinburgh.

[MACD97b] Scott McDonald, 1997. *A Context-based Model of semantic similarity.* Unpublished.

[MACD00] Scott McDonald, 2000. *Environmental determinants of lexical processing effort.* PhD dissertation, University of Edinburgh.

[MANN99] Christopher D. Manning and Hinrich Shütze, 1999. *Foundations of Statistical Natural Language Processing.* MIT Press.

[MARZ93] A. Marzal and E. Vidal,1993. *Computation of Normalized Edit Distance and Applications.* IEEE trans. on Pattern Analysis and Machine Intelligence, PAMI-15(9), 926-932.

[MAST61] Masterman, M, 1961. *Translation, Aristotelian Society Supp (context).*

[MILL91] George A. Miller and Walter G. Charles, 1991. *Contextual correlates of semantic similarity.* Language and Cognitive Processes, 6(1), 1991.

[OGDE23] C.K.Ogden and I.A. Richards, 1923. *The meaning of meaning.* Routledge & Kegan Paul LTD, Broadway House , 8th edition.

[PERE92] Pereira Fernando C.N., Naftali Z. Tishby, 1992. *Distributional Similarity, Phase Transitions and Hierarchical Clustering.* Working Notes, Fall Symposium Series, AAAI, pp. 108-112.

[PERE93] Pereira, Fernando C. N., Naftali Z. Tishby, and Lillian Lee, 1993. *Distributional clustering of English words.* In 30th Annual Meeting of the Association for Computational Linguistics: 183—190.

[RESN95] Philip Resnik, 1995. *Using information Content to evaluate semantic similarity in a Taxonomy.* Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995.

[RICH97] Stephen Richardson, 1997. *Determining similarity and inferring relations in a lexical knowledge base.* PhD. dissertation, City University of New York.

[ROSS76] Ross, Sheldon, 1976. *A First Course in Probability.* Macmillan.

[RUBE65] Herbert Rubenstein and John Goodenough, 1965. *Contextual correlates of synonymy.* Computational Linguistics, (8): 627-633.

[SADL89] V. Sadler 1989. *Working with analogical semantics: Disambiguation techniques in DLT.* Foris publications.

[SANK83] D. Sankoff and J.B. Kruskal, 1983. *Time Warps, String Edits and Macromolecules: The theory and practice of sequence comparison.* Addison-Wesley Reading, MA.

[SAUS16] Ferdinand de Saussure, 1916. *Cours de linguistique Generale.* Translated by Roy Harris as Course in General linguistics

[SCHU92] Hinrich Shütze, 1992. *Dimensions of meaning.* In Proceeding of Supercomputing '92,pp. 787-796, Los Alamitos, CA.

[SINC66] John McH, Sinclair, 1966. *Beginning the Study of Lexis.* In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins editors, In Memory of J. R. Firth, pages 410-430. Longmans, Green, and Co. Ltd,London.

[SOWA84] John F. Sowa, 1984. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley 1984.

[TVER77] A. Tversky, 1977. *Features of similarity.* Psychological Review, 84,327-352.

[WAGN74] R. A. Wagner and M. J. Fisher, 1974. *The string-to-string correction problem.* Journal of the Association of Computing Machinery, 21(1):168-173, January 1974.

[WATE94] Scott A. Waterman, 1994. *An associative Model of Word Use.* PhD Thesis, Brandeis University, computer science department.

[WIER72] WIERZBICKA Anna, 1972. *Semantic Primitives.* Frankfurt: Athenäum

[WITT72] Ludwig Wittgenstein, 1972. *Philosophical Investigations.*