Automatic Acquisition of Possible Contexts for Low-Frequent Words

Silvia Necsulescu
IULA
Universitat Pompeu Fabra
Barcelona, Spain
silvia.necsulescu@upf.edu

Abstract

The present work constitutes a PhD project that aims to overcome the problem caused by data sparsity in the task of acquisition of lexical resources. In any corpus of any length, many words are infrequent, thus they co-occur with a small set of words. Nevertheless, they can co-occur with many other words. Our goal is to discover some more possible co-occurring words for low-frequent words relying on other co-occurrences observed in corpus. Our approach aims to formulate a new similarity measure, based on the words usage in language, to approve a transfer of co-occurring words, from a frequent word to a low-frequent word.

1. Introduction

The production of language resources (LR) is a bottleneck for the development of many Natural Language Processing applications. The development of language resources by humans is very expensive and time consuming. Currently, a mainstream line of research is working on the automation of this task by using Machine Learning classifiers. To create language resources, first and foremost, automatic systems are needed to induce information from selected co-occurrences among words.

Any corpus is characterized by Zipf's law which states that the frequency of words is inversely proportional to its rank in the frequency table (Zipf, 1935). Words in a text follow a power-law distribution and many words show a low-frequency of occurrence, causing the problem known as data sparsity. Low-frequent words do not provide enough information for automatic systems that rely on the distributional information of a target word, i.e. co-occurrences with other words in a context (Bel, et al. 2007). Therefore, the frequency of words is a pitfall in the automatic production of LRs.

To overcome it, the low-frequent words need additional information to be classified by an automatic system. Bybee (2010) suggests that in order to process low frequent words, we can take evidence from other similar words. Thus we want to define *"similar words"*. For this task, a similarity measure implies to gather co-occurring words from frequent words to be used as virtual input of non frequent ones.

The word co-occurrences vary from one domain to another. We aim to create a generic system that takes into account the domain in an automatic manner. Therefore, to be able to identify suitable co-occurring words for a specific domain, we use a list of examples classified a priori, which is the only external knowledge provided.

The present article contains examples in Spanish and English to highlight that the problem of data sparsity exists in any language. We aim to create a language independent system, developed over a Spanish corpus and later, tested over an English corpus.

The rest of the paper is organized as follows: section 2 shows that low-frequent words represented a pitfall in previous works. In section 3, we introduce our objective and the main hypothesis that motivates this work, while in section 4 we present the proposed methodology. In section 5 we emphasis the contribution of our work and we formulate our conclusion over the present proposal.

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 121–126,
Hissar, Bulgaria, 13 September 2011.

2. Related Work

There have been different proposals on word similarity, for instance the (Frakes and Baeza-Yates., 1992), Jaccard's coefficient (Salton and McGill., 1983), Kullback-Leibler divergence measure (Kullback and Leibler, 1951), the L_1 norm (Kaufman, and Rousseeuw, 1990), Lin's measure (Lin, 1998), etc. Each of these measures is based on a description of the distributional behavior of each word in terms of other co-occurring words. To calculate the similarity between two words, the similarity between these vectors of co-occurrences is calculated.

These proposals, however, are not useful to handle words that occur just a few times in a corpus, as they do not give enough evidence on their distributional behavior. Therefore, although they are the most numerous set of words, most of the research done in various sub-tasks of the extraction of LR simply ignores low frequent words because the information provided is not enough to be reliable. For instance, Lin (1998) applies his measure of similarity on words that occurred at least 50 times in corpus. Rapp (2002) eliminated all words with a corpus frequency less than 101 to extract word associations from text. In the creation of language models, Padó and Lapata. (2003) removed infrequent words with occurrences less than 100. Peirsman, et al. (2008) considered as valid co-occurring words, only words that occurred at least 5 times.

In a general evaluation of various similarity measures for LR extraction, Curran and Moens (2002) eliminates all words with a frequency lower than 5, while Weeds and Weir (2005) consider the co-occurrences of a word with a frequency lower than 27 do not provide reliable information to describe its distributional behavior.

For our project, we aim to find more possible co-occurrences for words whose frequency is lower than 100. We face up to two problems, one is to extract the significant information for a low frequent word and the second one is to find a new measure of similarity that can handle the reduced information attached to low-frequent words.

Weeds and Weir (2005) tackle the problem of finding unseen co-occurrences of words by using the existent co-occurrences in corpus. As they rely on existing standard similarity measures and use as features, syntactic related words, they do not overcome the data sparsity problem.

3. Objective and Hypothesis

As previous work proved, any corpus of any size contains many low-frequent words, which do not provide enough information about their distributional behavior in language. Nevertheless, any word in human language can co-occur with a large set of words, while the co-occurrences in text represent just a small sub-set of this set.

Our objective is to overcome data sparsity by discovering other possible co-occurring words for low-frequent words besides the co-occurrences observed in corpus. In this way, we provide to low-frequent words, additional contextual information that allows them to be correctly handled in a further task.

To attain this objective, we rely on the distributional hypothesis (Harris, 1954), i.e. similar words tend to be used in similar contexts, and on Bybee's (1988,2010) statement that there is a similarity between a frequent word and a low-frequent word induced abstraction process over language. Bybee suggested that low frequent words can be processed by taking or copying information of more frequent similar words.

The challenge for our project is to discover a new topological space where we can define a measure of similarity based on distributional behavior of words that can handle low-frequent words. We propose a topology based on a graph representation of the lexicon.

Geffet and Dagan (2005) proved that although two words are similar in their distributional behavior, they do not share all co-occurrences. Hence, after we declare two words similar in usage, we must determine what words can be transferred from one word to another.

Therefore, our hypothesis is that relying on the representation of words in a graph that models relations among them, we can define a similarity measure that allows to calculate the probability of success for the transfer of co-occurring words, from a frequent word to a low-frequent one.

In the next sentence the word "entangled" occurs just 53 times in the British National Corpus.:

```
Some horses become excited
and upset if something goes
a bit wrong when they are in
harness, such as chains or
ropes becoming entangled
around their feet.
```

But, its context contains frequent word, such as *harness* (841), *chain* (5181), *rope* (2186) and *feet* (13349). More, the pattern "*become [...] around*" occur 15 times in corpus. In this pattern, in the slot where *entagled* occurs, we find also *destructive* (778), *apparent* (5216), *wrapped* (1613), *unstable* (697), *millstones* (102), *known* (25176), *mobilized* (122), *centered* (31), *noticeable* (826), *compacted* (84), *deadlocked* (67). We suppose that some of these words are similar in their usage with the target word "*entagled*" and between their contexts we can find possible co-occurrences for the word "*entagled*"

4. Methodology

In an initial step, we aim to model IULA Spanish Corpus (Cabré et al. 2006) in a graph structure, to shed light over the relations that exist between words influenced by their context or by their lexical-morphological features.

Next, using the language topology created before, we aim to define a new measure of similarity between words to associate a low-frequent word with a frequent one, plausible for a transfer of co-occurring words.

Finally, a probabilistic model is created to calculate the probability of two words, unseen before together in context, to co-occur together.

4.1 Graph Model

To create the graph language model, we represent the corpus lexicon in nodes and there is an edge between two nodes, if they are contextual related or similar at lexical-morphological structure.

The contextual relations between words in our language topology are resulted from both *syntagmatic relations*, i.e. words that co-occur in the same context in the same time more frequently than expected by chance and *paradigmatic relations*, i.e. words that occur in the same context, but not in the same time. Thus, we will take advantage of all the information available.

Syntagmatic related words are the co-occurrences seen in corpus. The most key part in the graph design is to set up those syntagmatic relations that provide us with reliable information for low-frequent words i.e. words that co-occur in the same context and which manifest lexical-semantic affinities beyond grammatical restrictions (Halliday, 1966).

There are two mainstream lines to define syntagmatic related words: focused on the proximity in text or on the syntactic relations between them.

Besides, and differently to other authors, because we have very little information, we take into account all determiners and modifiers. The position in an area of text is not a strong enough constraint to extract exactly those words that are significant. For instance, word's modifiers or determinants can be outside of a fixed area of text while in the word proximity we can find useless information.

Meanwhile, to extract co-occurrences defined by syntactical relationships, a parser is needed to be applied. Nevertheless, the use of a parser has some drawbacks, such as a large preprocessing step and sparse information extracted. Therefore, for the extraction of the syntagmatic relations reliable for low-frequent words, we define heuristic rules, stronger than the simple presence in an area of text and looser than syntactic relations.

Ferrer i Cancho and Solé, (2001) stated that the most significant part of co-occurrences in sentence is due to syntactical relationships between words, e.g. head-modifier or dependency relationships, but also due to stereotyped expressions or collocations, e.g. take it easy, New York. More, Ferrer i Cancho et al. (2007) assumed the importance of the frequencies of word co-occurrences, while Choudhury et al. (2010) suggested the importance of the part-of-speech category in the language organization.

To define the syntagmatic relations, we aim to find statistical information that characterizes words syntactic related. We extract statistical information from corpus about word frequency, part-of-speech and co-occurring words in the same paragraph and we apply a parser. Finally, we mix the statistical information with the syntactic relationships, to formulate heuristic rules to be applied over raw text with the goal to extract those co-occurring words that are syntagmatic related with a target word.

Using reliable syntagmatic relations defined, we calculate paradigmatic relations to discover words that share the same context but in different moments. To determine paradigmatic related words, we compare their co-occurrences vector using one of the standard similarity measure, e.g. Lin's measure (Lin, 1998).

Syntagmatic or paradigmatic relations represent the syntactic behavior of a word. In human language, the interaction of words in an utterance is not separated by their lexical-morphological structure, e.g. in English, the verb *avoid* must be followed by a verb at *–ing* form. Therefore, we add in the graph structure an edge between words that are similar from the point of view of their lexical-morphological features, i.e. they present the same affixes or the same root.

4.2 The Structure Analysis

The graph model created previously represents the language topology. It contains linguistic relations, created from two points of view, first, relations that represent the combination of words in sentences and second, relations that connect similar words regarding their lexical-morphologic features. Relying on this topology, the next step is to define the measure of similarity between two words for a possible transfer of co-occurrences between them.

The previous studies over various models of language give us the intuition of the existence of common patterns in the large scale language organization. Language models created with co-occurrences (Ferrer i Cancho, et al., 2001) and syntactic relations (Ferrer i Cancho, et al., 2007) followed the same pattern of complex networks, characterized by a *small-world* structure (Watts, 1999) and a *scale-free* structure (Barabási, et al., 1999). The former presents a small average path length between vertices, a sparse connectivity (i.e. a node is connected to only a very small percentage of other nodes), and a strong local-clustering (i.e. the extent to which the neighborhoods of neighboring nodes overlap). The latter means that the number of vertices with degree *k* falls off as an inverse power of *k*, consequently, the majority of words have relatively few connections joined together through a small number of hubs with many connections.

To calculate the measure of similarity, first we want to extract general information over the graph structure, such as the type of words that are hubs and the type of words that are related with them, common properties of these words or what clusters of words are created and the common properties of them. Because, in our model, we use syntagmatic relations created in a heuristic manner, paradigmatic relations, and also similarity relations extracted from the internal word structure, first and foremost we have to verify if our topology keeps the complex network structure.

After we extracted the general information, to formulate the similarity measure, we focus on two axes. On one hand, we create clusters of words that occur in the same slot of a language pattern and we search similarities between the words structure from the same cluster (Bybee, 2006). On the other hand, we provide a list of words a priori classified and we search statistical similarities between the structures of words from the same class. To be able to analyze the importance of various structural features over the measure of similarity, such as the number of connections, the connection types or the connections with various classes of words, we search a response for the next questions:

- What are the features that characterize each word?
- What types of words are connected in topology with our words?
- What features have the structure that links two words from the same class/cluster?
- What is common in the structures of all the words from a class/cluster?

For a short illustration of our procedure, we created a graph using as corpus the sentences listed below, extracted from the IULA Spanish Corpus (Cabré, et al., 2006). The heuristic rule used to extract the syntagmatic related words is "*<noun> potential*". All the words that occur in the slot *<noun>* are related by a paradigmatic relation. For a better understanding of the graph, we do not draw these relations. The nouns ending with the suffix *–nte* are inter-connected with an edge for their lexical-morphological similarity. Analyzing the created graph, we observed that all nouns ended with the suffix *–nte* represent *human beings* and they are clustered by the lexical-morphological edges.

Carga Q se define como ## la energía potencial ## que posee una carga q […]

[…] juicio de los analistas, ## el precio potencial ## de la sociedad en un […]

El Consejo de Lisboa incrementó ## el crecimiento potencial ## de nuestras economías.

[…] preservar, siquiera mínimamente, ## el riesgo potencial ## , pero cierto , de […]

Las inversiones son altas , unos 300 millones de dólares , porque ## los clientes potenciales ## se estiman en 20 millones .

Los demócratas intentan asustar a ## los votantes potenciales ## de Nader asegurando que Bush pondría en peligro

Las nuevas reglamentaciones exigen examinar a ## los donantes potenciales ## de todo tipo de tejidos.

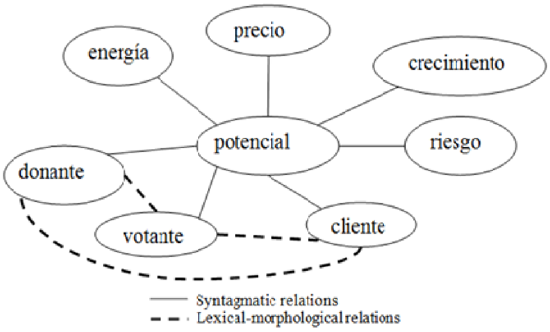

Figure 1: The graph language model created with the previous examples

The co-occurrences of a word are dependent on the domain. Therefore, we harvest on the one hand general features of the structure of the same corpus by comparing words that are used in the same language pattern and on the other hand, domain related co-occurrences by comparing words from the same class in that domain. By combining these two results, we define a measure of similarity appropriate for the given domain, wherever it is the general domain or a specialized one, and focused on the type of lexical resources that aim to be produced further.

4.3 The Probabilistic Model

Using the results of the previous stages, the graph language model and the similarity measure defined using the graph model, we create a probabilistic model.

We aim to calculate the probability that a target word *w* can occur with another word *f*, existent in corpus, in an utterance, even if this co-occurrence is not seen in context.

To calculate this probability we rely on each word similar in the topological model with *w*. We search between its co-occurring words f_i, a word that is similar with *f*. The final probability depends on the similarity between the word *w* and w_i, the similarity between *f* and f_{ij} and the probability that w_i occurs with f_{ij}. $P(f_{ij}|w_i)$ is 1 if this co-occurrence is seen in corpus. The next formula is the mathematical expression used to calculate the probability of co-occurrence.

$$P(f|w) = \frac{\sum_{i=1}^{|V(w)|}\sum_{j=1}^{|F(w_i)|} Sim(w,w_i)Sim(f,f_{ij})P(f_{ij}|w_i)}{\sum_f \sum_{i=1}^{|V(w)|}\sum_{j=1}^{|F(w_i)|} Sim(w,w_i)Sim(f,f_{ij})P(f_{ij}|w_i)}$$

Where

- w_i is a connected word with *w*
- *f* is a co-occurring word with *w*
- f_{ij} is a co-occurring word with w_i
- *V(w)* is the set of similar words to w calculated using the relations from graph
- *F(w_i)* is the set of co-occurring words with wi
- *Sim(x,y)* is the similarity measure defined previously that calculates the similarity between the word *x* and *y*

Using the probabilistic model we decide which co-occurring words are transferred from one word to another. We provide, for the low-frequent words new possible co-occurring words. As a consequence, the context of low-frequent words is larger and therefore, they can be further classified by a system of automatic acquisition of lexical resources.

5. Contributions of the Work and Conclusions

The word frequency in a corpus is a bottleneck in the automatic acquisition of LRs based on corpus. In any corpus, there are many words whose context does not provide enough information to classify them. Our approach is based on the combination of words in valid utterances, to find a solution to overcome the data sparsity.

The importance of the work relies on our focus on low frequent words. As we showed previously, in different task of corpus analysis, a cutoff was applied over the words frequency to eliminate those words whose contextual information was small and consequently, not reliable. We aim to develop a new similarity measure, focused on low-frequent words, which differently than other standard measure of similarity is based on the graph model. This model contains edges that relate words from the same context, words that share the same context but in different moments and also words with a similar lexical-morphological structure.

Differently to previous work, our measure of similarity does not imply a semantic similarity, but a similarity at the distributional behavior that allows a transfer of co-occurring words from the most frequent word to the less frequent one.

If our hypothesis is valid, relying on the language topology created with various relation types, we induce more likely co-occurrences for low-frequent words. Further, our results can be used for the automatic acquisition of lexical resources to cover different domains and different languages.

Acknowledgments
I want to thank Nuria Bel and Muntsa Padró for their valuable indications in the development of the subject. This project is funded by the CLARA project (EU-7FP-ITN-238405).

References

Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. Science , 286.

Bel, N., Espeja, S., and Marimon, M. (2007). Automatic Acquisition of Grammatical Types for Nouns. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics;.

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. Language 82(4). , pp. 711-733.

Bybee, J. L. (1988). Morphology as lexical organization. In M. H. Noonan, Theoretical morphology (pp. 119-141.). Academic Press.

Bybee, J. (2010). Language, usage and cognition.

Cabré, M. T., Bach, C., and Vivaldi, J. (2006). 10 anys del Corpus de l'IULA. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Choudhury, M., Chatterjee, D., and Mukherjee, A. (2010). Global topology of word co-occurrence networks: Beyond the two-regime power-law. proceedings of COLING(10). Beijing, China.

Curran, J. R., and Moens, M. (2002). Improvements in automatic thesaurus extraction. PROCEEDINGS OF THE WORKSHOP ON UNSUPERVISED LEXICAL ACQUISITION .

Ferrer i Cancho, R., and Solé, R. (2001). The small-world of human language. Proceedings of the Royal Society of London .

Ferrer i Cancho, R., Mehler, A., Pustylnikov, O., and Díaz-Guilera, A. (2007). Correlations in the organization of large-scale syntactic dependency networks.

Frakes, W. B., and Baeza-Yates, R. (1992). Information Retrieval, Data. Prentice Hall.

Geffet, M., and Dagan, I. (2005). The Distributional Inclusion Hypotheses and Lexical Entailment. The 43rd Annual Meeting of the Association for Computational Linguistics.

Halliday, M. (1966). Lexis as a linguistic level. In memory of JR Firth , pp. 148–162.

Harris, Z. S. (1954). Distributional structure.

Kaufman, L., and Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: JohnWiley.

Kullback, S., and Leibler., R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics .

Lin, D. (1998). An information-theoretic definition of similarity. Proceedings of International Conference on Machine Learning. WI: Madison.

Padó, S., and Lapata, M. (2003). Constructing Semantic Space Models from Parsed Corpora. IN PROCEEDINGS OF ACL-03, (pp. 128--135).

Peirsman, Y., Heylen, K., and Speelman, D. (2008). Putting things in order. First and second order context models for the calculation of semantic similarity. Actes des 9es Journées internationales d'Analyse statistique des Données textuelles (JADT 2008). Lyon: France.

Rapp, R. (2002). The Computation Of Word Associations: Comparing Syntagmatic And Paradigmatic Approaches. coling.

Salton, G., and McGill, M. J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill: New York.

Watts, D. J. (1999). Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton University Press.

Weeds, J., and Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. Computational Linguistics , 31.

Zipf, G. K. (1935). The Psychobiology of Language. Houghton-Mifflin.