Corpus representativeness for syntactic information acquisition

Núria BEL
IULA, Universitat Pompeu Fabra
La Rambla 30-32
08002 Barcelona
Spain
nuria.bel@upf.edu

Abstract

This paper refers to part of our research in the area of automatic acquisition of computational lexicon information from corpus. The present paper reports the ongoing research on corpus representativeness. For the task of inducing information out of text, we wanted to fix a certain degree of confidence on the size and composition of the collection of documents to be observed. The results show that it is possible to work with a relatively small corpus of texts if it is tuned to a particular domain. Even more, it seems that a small tuned corpus will be more informative for real parsing than a general corpus.

1 Introduction

The coverage of the computational lexicon used in deep Natural Language Processing (NLP) is crucial for parsing success. But rather frequently, the absence of particular entries or the fact that the information encoded for these does not cover very specific syntactic contexts --as those found in technical texts— make high informative grammars not suitable for real applications. Moreover, this poses a real problem when porting a particular application from domain to domain, as the lexicon has to be re-encoded in the light of the new domain. In fact, in order to minimize ambiguities and possible over-generation, application based lexicons tend to be tuned for every specific domain addressed by a particular application. Tuning of lexicons to different domains is really a delaying factor in the deployment of NLP applications, as it raises its costs, not only in terms of money, but also, and crucially, in terms of time.

A desirable solution would be a 'plug and play' system that, given a collection of documents supplied by the customer, could induce a tuned lexicon. By 'tuned' we mean full coverage both in terms of: 1) entries: detecting new items and assigning them a syntactic behavior pattern; and 2) syntactic behavior pattern: adapting the encoding of entries to the observations of the corpus, so as to assign a class that accounts for the occurrences of this particular word in that particular corpus. The question we have addressed here is to define the size and composition of the corpus we would need in order to get necessary and sufficient information for Machine Learning techniques to induce that type of information.

Representativeness of a corpus is a topic largely dealt with, especially in corpus linguistics. One of the standard references is Biber (1993) where the author offers guidelines for corpus design to characterize a language. The size and composition of the corpus to be observed has also been studied by general statistical NLP (Lauer 1995), and in relation with automatic acquisition methods (Zernick, 1991, Yang & Song 1999). But most of these studies focused in having a corpus that actually models the whole language. However, we will see in section 3 that for inducing information for parsing we might want to model just a particular subset of a language, the one that corresponds to the texts that a particular application is going to parse. Thus, the research we report about here refers to aspects related to the quantity and optimal composition of a corpus that will be used for inducing syntactic information.

In what follows, we first will briefly describe the observation corpus. In section 3, we introduce the phenomena observed and the way we got an objective measure. In Section 4, we report on experiments done in order to check the validity of this measure in relation with word frequency. In section 5 we address the issue of corpus size and how it affects this measure.

2 Experimental corpus description

We have used a corpus of technical specialized texts, the CT. The CT is made of subcorpora belonging to 5 different areas or domains: Medicine, Computing, Law, Economy, Environmental sciences and what is called a General subcorpus made basically of news. The size of the subcorpora range between 1 and 3 million words per domain. The CT corpus covers 3 different languages although for the time being we

have only worked on Spanish. For Spanish, the size of the subcorpora is stated in Table 1. All texts have been processed and are annotated with morphosyntactic information.

The CT corpus has been compiled as a test-bed for studying linguistic differences between general language and specialized texts. Nevertheless, for our purposes, we only considered it as documents that represent the language used in particular knowledge domains. In fact, we use them to simulate the scenario where a user supplies a collection of documents with no specific sampling methodology behind.

3 Measuring syntactic behavior: the case of adjectives

We shall first motivate the statement that parsing lexicons require tuning for a full coverage of a particular domain. We use the term "full coverage" to describe the ideal case where we would have correct information for all the words used in the (unknown a priori) set of texts we want a NLP application to handle. Note that full coverage implies two aspects. First, type coverage: all words that are used in a particular domain are in the lexicon. Second, that the information contained in the lexicon is the information needed by the grammar to parse every word occurrence as intended.

Full coverage is not guaranteed by working with 'general language' dictionaries. Grammar developers know that the lexicon must be tuned to the application's domain, because general language dictionaries either contain too much information, causing overgeneration, or do not cover every possible syntactic context, some of them because they are specific of a particular domain. The key point for us was to see whether texts belonging to a domain justify this practice.

In order to obtain objective data about the differences among domains that motivate lexicon tuning, we have carried out an experiment to study the syntactic behavior (syntactic contexts) of a list of about 300 adjectives in technical texts of four different domains. We have chosen adjectives because their syntactic behavior is easy to be captured by bigrams, as we will see below. Nevertheless, the same methodology could have been applied to other open categories.

The first part of the experiment consisted of computing different contexts for adjectives occurring in texts belonging to 4 different domains. We wanted to find out how significant could different uses be; that is, different syntactic contexts for the same word depending on the domain. We took different parameters to characterize what we call 'syntactic behavior'.

For adjectives, we defined 5 different parameters that were considered to be directly related with syntactic patterns. These were the following contexts: 1) pre-nominal position, e.g. 'importante decisión' (*important decision*) 2) post-nominal position, e.g. 'decisión importante' 3) 'ser' copula[1] predicative position, e.g. 'la decisión es importante' (*the decision is important*) 4) 'estar' copula predicative position, e.g. 'la decisión está interesante/*importante' (*the decision is interesting/important*) 5) modified by a quantity adverb, e.g. 'muy interesante' (*very interesting*). Table 1 shows the data gathered for the adjective "paralelo" (*parallel*) in the 4 different domain subcorpora. Note the differences in the position 3 ('ser' copula) when observed in texts on computing, versus the other domains.

Corpora/n.of occurrences	1	2	3	4	5
general (3.1 M words)	1	61	29	3	0
computing (1.2 M words)	4	30	0	0	0
medecine (3.7 M words)	3	67	22	1	0
economy (1 M words)	0	28	6	0	0

Table 1: Computing syntactic contexts as behaviour

The observed occurrences (as in Table 1) were used as parameters for building a vector for every lemma for each subcorpus. We used *cosine distance*[2] (CD) to measure differences among the occurrences in different subcorpora. The closer to 0, the more significantly different, the closer to 1, the more similar in their syntactic behavior in a particular subcorpus with respect to the general subcorpus. Thus, the CD values for the case of 'paralelo' seen in Table 1 are the following:

Corpus	Cosine Distance
computing	0.7920
economy	0.9782
medecine	0.9791

Table 2: CD for 'paralelo' compared to the general corpus

[1] Copulative sentences are made of 2 different basic copulative verbs 'ser' and 'estar'. Most authors tend to express as 'lexical idyosincracy' preferences shown by particular adjectives as to go with one of them or even with both although with different meaning.

[2] Cosine distance shows divergences that have to do with large differences in quantity between parameters in the same position, whether small quantities spread along the different parameters does not compute significantly. Cosine distance was also considered to be interesting because it computes relative weight of parameters within the vector. Thus we are not obliged to take into account relative frequency, which is actually different according to the different domains.

What we were interested in was identifying significant divergences, like, in this case, the complete absence of predicative use of the adjective 'paralelo' in the computing corpus. The CD measure has been sensible to the fact that no predicative use has been observed in texts on computing, the CD going down to 0.7. Cosine distance takes into account significant distances among the proportionality of the quantities in the different features of the vector. Hence we decided to use CD to measure the divergence in syntactic behavior of the observed adjectives. Figure 1 plots CD for the 4 subcorpora (Medicine, Computing, Economy) compared each one with the general subcorpus. It corresponds to the observations for about 300 adjectives, which were present in all the corpora. More than a half for each corpus is in fact below the 0.9 of similarity. Recall also that this mark holds for the different corpora, independently of the number of tokens (Economy is made of 1 million words and Medicine of 3).

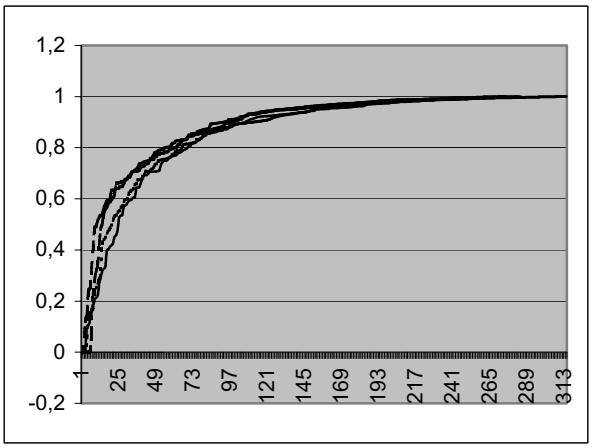

Figure 1: Cosine distance for the 4 different subcorpus

The data of figure 1 would allow us to conclude that for lexicon tuning, the sample has to be rich in domain dependent texts.

4 Frequency and CD measure

For being sure that CD was a good measure, we checked to what extent what we called syntactic behavior differences measured by a low CD could be due to a different number of occurrences in each of the observed subcorpora. It would have been reasonable to think that when something is seen more times, more different contexts can be observed, while when something is seen only a few times, variations are not that significant.

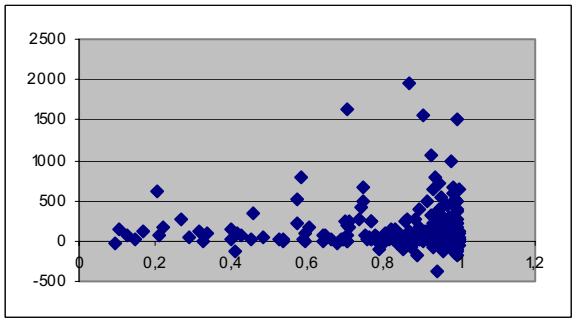

Figure 2: Difference in n. of observations in 2 corpora and CD

Figure 2 relates the obtained CD and the frequency for every adjective. For being able to do it, we took the difference of occurrences in two subcorpora as the frequency measure, that is, the number resulting of subtracting the occurrences in the computing subcorpus from the number of occurrences in the general subcorpus. It clearly shows that there is no regular relation between different number of occurrences in the two corpora and the observed divergence in syntactic behavior. Those elements that have a higher CD (0.9) range over all ranking positions: those that are 100 times more frequent in one than in other, etc. Thus we can conclude that CD do capture syntactic behavior differences that are not motivated by frequency related issues.

5 Corpus size and syntactic behavior

We also wanted to see the minimum corpus size for observing syntactic behavior differences clearly. The idea behind was to measure when CD gets stable, that is, independent of the number of occurrences observed. This measure would help us in deciding the minimum corpus size we need to have a reasonable representation for our induced lexicon. In fact our departure point was to check whether syntactic behavior could be compared with the figures related to number of types (lemmas) and number of tokens in a corpus. Biber 1993, Sánchez and Cantos, 1998, demonstrate that the number of new types does not increase proportionally to the number of words once a certain quantity of texts has been observed.

In our experiment, we split the computing corpus in 3 sets of 150K, 350K and 600K words in order to compare the CD's obtained. In Figure 3, 1 represents the whole computing corpus of 1,200K for the set of 300 adjectives we had worked with before.

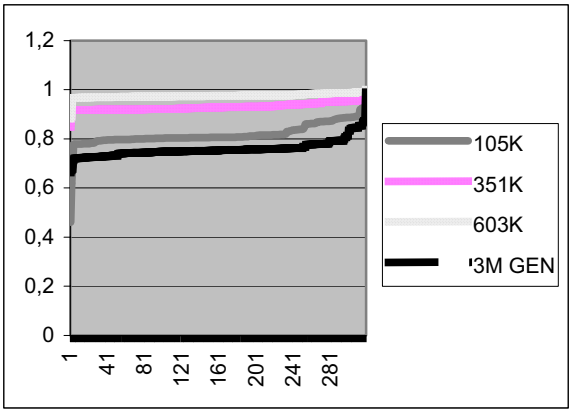

Figure 3: CD of 300 adjs. in different size subcorpora and general corpus

As shown in Figure 3, the results of this comparison were conclusive: for the computing corpus, with half of the corpus, that is around 600K, we already have a good representation of the whole corpus. The CD being superior to 0.9 for all adjectives (mean is 0.97 and 0.009 of standard deviation). Surprisingly, the CD of the general corpus, the one that is made of 3 million words of news, is lower than the CD achieved for the smallest computing subcorpus. Table 3 shows the mean and standard deviation for all de subcorpora (CC is Computing Corpus).

Corpus	size	mean	st. deviation
CC	150K	0.81	0.04
CC	360K	0.93	0.01
CC	600K	0.97	0.009
CC	1.2 M	1	0
General	3M	0.75	0.03

Table 3: Comparing corpus size and CD

What Table 3 suggests is that according to CD, measured as shown here, the corpus to be used for inducing information about syntactic behavior does not need to be very large, but made of texts representative of a particular domain. It is part of our future work to confirm that Machine Learning Techniques can really induce syntactic information from such a corpus.

References

Biber, D. 1993. Representativeness in corpus design. Literary and Linguistic Computing 8: 243-257.

Lauer, M. 1995. "How much is enough? Data requirements for Statistical NLP". In 2nd. Conference of the Pacific Association for Computational Linguistics. Brisbane, Australia.

Sánchez, A. & Cantos P., 1997, "Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora, A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish," *In International Journal of Corpus Linguistics* Vol. 2, No. 2.

Schone, P & D. Jurafsky. 2001. Language-Independent induction of part of speech class labels using only language universals. Proceedings IJCAI, 2001.

Yang, D-H and M. Song. 1999. "The Estimate of the Corpus Size for Solving Data Sparseness". Journal of KISS, 26(4): 568-583.

Zernik, U. Lexical Acquisition. 1991. Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum Associates: 1-26.