Collocation Extraction and Text Analysis: Different Types of Collocations and Different Genres

Lidia Pivovarova, Elena Yagunova
Saint-Petersburg State University
lidia.pivovarova@gmail.com, iagounova.elena@gmail.com

Introduction. Methodology. Hypotheses

We take ***collocation*** to mean any nonrandom cooccurrence of two or more lexical units, specific for either language as a whole or particular genre of texts (corpus). There are many papers aimed at finding out collocations for investigation of language specific units,: idioms, or compound functional words, or compound names, etc. However, the style of corpora texts influences the list of extracted collocations as well as the appropriate collocation extraction techniques (Ferret et al. 1998). We assume that collocations can represent formal markers of genre, subject domain and stylistic features of the corpus.

For different corpora, lists of probable collocations have to be open, as a substantial part of collocations features cannot be set *a priori*. Thus the statistical approach seems to be the most appropriate in our case. However, any statistical measure has its own nature. We use two well-known measures: MI and t-score Church et al., 1991; Stubbs, 1995). For both measures, we use lemma-collocations and wordform-collocations (since Russian has a very large morphology paradigm).

Our material is two Russian corpora which represent two genre of texts): i) ***News corpus*** – texts from news portal www.lenta.ru for 2009 year whose size is approximately 10 millions tokens (words and punctuation marks) and ii) ***Scientific corpus*** – Proceedings of International St. Petersburg conference “Corpus Linguistics” for 2004-2008 year, the size of the Russian part of this corpus is approximately 250 thousands tokens. Both corpora have been automatically lemmatized.

The first measure used, Mutual Information (MI), “can be understood as a coefficient of association strength” (Evert, 2005). MI for a digram is given by the formula (Church, 1991):

$$MI = log2 \frac{f(n,c) \times N}{f(n) \times f(c)}$$

, where the meaning of notations is as follows:

MI , mutual information;

f(n,c), f(n), f(c), absolute frequencies of occurence of digram xy and words x, y respectively, N, the collection size.

According to our experiments, the numerical value of MI depends on the corpus size. However is very difficult to collect a very large homogeneous scientific corpus (with a restricted subject domain); this is why our corpora have different size. In this paper we ignore the MI numerical values and use only rank order. We analyze the top 100 or the top 1000 collocations ordered by MI.

Another disadvantage of MI is overestimating of low-frequency collocations (Stubbs, 1995); this is why some cutoff has to be used. We use a threshold of 40 for “Lenta.ru” and 16 for “Corpus linguistics” corpora below. These thresholds were set according to our interest in a subject domain – after comparing results with different cutoff frequencies (we are going to address the formal methods for this cutoff problem in the nearest future).

The second measure, t-score, can be understood as a modification of the collocation frequency, the built-in correction “has a large effect only with a small number of common grammatical words” (Stubbs (1995)). T-score for digram is given by the formula (Stubbs, 1995):

$$t-score = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}}$$

We analyze the top 100 or the top 1000 collocations ordered by t-score, thus all comparable lists of collocations have the same size.

The main hypotheses underlying this paper are: (1) automatically extracted lists of collocations can represent formal markers of genre, subject domain and stylistic features of the corpus; (2) MI-collocations consists of such multiword expressions (MWE) as terminology (especially for scientific corpora) and nominations (organizations, persons, locations), this measure is usable for determining subject domain and genre; (3) t-score picks out functional, grammatical compounds and high-frequency constructions; (4) t-score is very useful for extracting MWEs which present in ***every*** (or almost every) text of collection; this feature is the most important for homogeneous corpora.

News corpus

The majority of the first 100 digrams ordered by MI for the news corpus are object nominations and proper names: persons (e.g. БРИТНИ СПИРС *Britney Spears*), organizations (e.g. ЛЕ БУРЖЕ *Le Bourget*) and locations (e.g. МЫС КАНАВЕРАЛ *Cape Canaveral*).

25% of digrams may be considered as terminology, news clichés, bureaucratisms: e.g. СЕРДЕЧНЫЙ ПРИСТУП (*heart attack* - term), СТИХИЙНЫЙ БЕДСТВИЕ (*natural disaster* - cliche), ТРОТИЛОВЫЙ ЭКВИВАЛЕНТ (*TNT equivalent* - bureaucratism).

Lists of MI-news-digrams show such features of genre as extension of terminology functions and a high amount of object nominations and proper names. Also, many digrams are very special for 2009: for example, *Nevsky Express*, the well-known name of a train route that was subject to a terrorist attack, or *hadron collider* which was started that year. However, we cannot say anything about the quality of topic identification by MI-score – it is necessary to carry out some additional experiments.

It is much easier to interpret the top set of the digrams ordered by t-score. This measure picks out high-frequency collocations. Many t-score-news-digrams are indications of **information sources** (e.g. ПО СЛОВАМ *according to*, РИА НОВОСТЬ *RIA News,* СО ССЛЫКА and ССЫЛКА НА might be fragments of СО ССЫЛКОЙ НА (*with reference to*). Very often these digrams are fragments of longer collocations. These collocations seem to be uninformative with respect to subject domain, but they are useful for detecting information sources.

The t-score picks out many grammatical compounds and high-frequency (for the news corpus) PREPOSITION+NOUN constructions: e.g. В ТЕЧЕНИЕ *during*, ВО ВРЕМЯ *throughout*, В РОССИЯ *in Russia*. The majority of these compounds are vocabulary units, but their list gives sufficient information about the genre and style of texts.

Scientific corpus

The top of bigrams ordered by MI for our homogeneous scientific corpus consists of both terminology and discursive expressions. However, a simple part-of-speech filter allows the separation of terms from other collocations, since most of them are noun groups: “речевой деятельности” (*speech perception and production*), “художественной литературы” (*fiction*), etc. The majority of other MI-collocations are discursive words: “наш взгляд” (*our point of view*), “крайней мере” (*at least*) etc.

The top of bigrams ordered by t-score for our scientific corpus are grammatical compounds, high-frequency PREPOSITION+NOUN constructions typical for this corpus (e.g., “таким образом” *in such a way*, “в качестве” *as*, “в виде” *in the form of*, “в корпусе” *in the corpus*), most of compounds are presented in Russian dictionaries, but they are more frequent in scientific texts.

T-score extracts terminological bigrams which are present in ***every*** (or almost every) texts of collection: КОРПУС ТЕКСТ (*text corpus*), ЧАСТЬ РЕЧЬ (*part of speech*), ЛЕКСИЧЕСКИЙ ЕДИНИЦА (*lexical unit*), МАШИННЫЙ ПЕРЕВОД (*machine translation*). A simple part-of-speech filter allows extracting cross-corpus terms with high accuracy. This way of term extracting might be useful to detect the level of homogeneity, especially for new and cross-discipline scientific domains where terminology is unsettled.

Conclusion

Our results prove the main hypotheses. MI picks out proper names, terms, object nominations, while t-score is much better in finding cross-language collocations (functional words, discursive words) and fixed constructions which characterize the corpus style. Thus both measures may be useful in determining the corpus genre and subject. For news collection, MI is much better for a subject, while t-score is better for a genre variability. For homogeneous scientific collections, the lists of collocations picked by MI and t-score have more intersections than for news collection since similar terminological collocations are present in every (almost every) paper in the collection and therefore these terms are picked both by t-score and MI measures.

In our report we also discuss difference in bigram lists for lemmas vs. those for wordforms. This is a question aimed at investigation of regularity of syntactic roles, especially for MI-bigrams. Usually the most informative MI-bigrams are placed at the intersection of lemmas and wordforms lists. Also we shall present some preliminary data from our ongoing research based on two other news corpora (http://www.ng.ru/ & http://www.compulenta.ru/) and one scientific corpus (Papers from the Annual International Conference "Dialogue" (2003-2009)).

The next stages of our research will probably include: experiments with informants to check our results; further extension of MI for n-gram of any length; comparison of MI and t-score with other statistical measures.

References

Ferret, O., Grau, B. and Masson, N. 1998. Thematic segmentation of texts: two methods for two kinds of texts. In Proceeding of Coling-ACL-1998, v. 1, pp. 392-396.

Church, K., Gale, W., Hanks, P. and Hindle, D. 1991. Using Statistics in Lexical Analysis. Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. New Jersey, Lawrence Erlbaum pp. 115-164.

Evert, S. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, University of Stuttgart.

Stubbs, M. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. Functions of Language, 1 pp 23–55.