Domain Collocation Identification

Jiří Materna

Centre for Natural Language Processing
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech Republic
xmaterna@fi.muni.cz
http://nlp.fi.muni.cz

Abstract. In this paper we present a new method of automatic collocation identification. Collocation is an important relation between words, which is widely used, among others, in information retrieval tasks. Over the last years, many methods of automatic collocation acquisition from text corpora have been proposed. The approach described in this paper differs from the others in focusing on domain collocations. By the domain collocation we mean a collocation which is specific for a relatively small set of documents related to the same topic. The proposed method has been implemented and used in a real information retrieval system. Comparing to the common non-domain approach, the precision of the system has increased significantly.

Key words: collocation; domain; information retrieval

1 Introduction

Lexical collocations are an important phenomenon in many natural language processing tasks like computational lexicography [1], word sense disambiguation [2], machine translation [3], information extraction [4], etc. In this work we focus on their exploitation in information retrieval. In our information retrieval system, there is a need for identifying collocations in queries in order to treat them as single units.

Let's have a look at a simple query *finanční úřad v Karlových Varech (tax office in Karlovy Vary)*. Combinatorially, there are many ways how to parse it but, in our point of view, the correct one is only *(finanční úřad) v (Karlových Varech)*. It means that the identified collocation (in this case *finanční úřad* and *Karlových Varech*) should not be split.

Collocation is an expression consisting of two or more associated words or tokens. Unfortunately, there is no formal linguistic definition. For us, the collocation is understood as an n-gram of tokens whose co-occurrence in a large text corpus is statistically outstanding. There are many statistical measures usable to detect collocations in corpora. Most of them are based on classical mathematical statistics (t-score, chi-square) or information theory (mutual information). All these methods are widely used and explored in many applications but they suffer from the following disadvantages:

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 17–21, 2009.

- The association scores are strongly influenced by the size of the corpus. Thus, the score values acquired from different corpora are not comparable and even the maximum or minimum values of the score may be different.
- The association scores are suitable for identifying global collocations but they are not convenient for domain-specific collocations, which are relevant only for a small set of documents related to the same topic.

The presented paper deals mainly with the second problem and is structured as follows. In the next section we will introduce the most commonly used statistical methods of collocation identification in text corpora. In Sectioni 3 we will focus on a new method of domain collocation acquisition. In this section we will also discuss advantages and disadvantages of the proposed method.

2 Statistical Approaches to Collocation Identification

Over the last years, many methods of automatic collocation acquisition from large text corpora have been proposed. All association scores which are subjects of this research use only word frequency characteristics. The simplicity and the ease of use belong to the most important advantages of the statistical approach. In order to ensure maximum readability of the paper we will only consider collocations consisting of two tokens but the described methods are universal. In the rest of the paper the following notation will be used:

- $f(t)$ – number of occurrences of term t in the whole corpus;
- $f(t_1, t_2)$ – number of co-occurrences of terms t_1, t_2 (by the co-occurrence we mean that the tokens occur in the corpus directly one after another);
- n – number of all tokens in the corpus.

T-score

This measure uses classical statistic approach based on Student's t-test [5]. The association score is defined as:

$$T(t_1, t_2) = \frac{f(t_1, t_2) - \frac{f(t_1)f(t_2)}{n}}{\sqrt{f(t_1, t_2)}}$$

MI-score

This measure comes from information theory and corresponds to the quantity of information given by the occurrence of one term about occurrences of another one. The mutual information association score is defined as:

$$MI(t_1, t_2) = \log_2 \frac{f(t_1, t_2)n}{f(t_1)f(t_2)}$$

MI^2-score

Mutual information score is a useful measure but it is strongly influenced by the frequency of tokens. To reduce this disadvantage same heuristics have been proposed [6]. One of the most popular is MI^2-score:

$$MI^2(t_1, t_2) = \log_2 \frac{f(t_1, t_2)^2 n}{f(t_1) f(t_2)}$$

Dice score

Dice score identifies pairs with a particularly high degree of lexical cohesion (i.e. those with nearly total association) [7]:

$$D(t_1, t_2) = \frac{2f(t_1, t_2)}{t_1 + t_2}$$

logDice score

Dice score gives a good association score but the problem is that the values are usually very small numbers. This is solved by the logDice score [8]:

$$logD = 14 + \log_2 \frac{2f(t_1, t_2)}{t_1 + t_2}$$

In many application we need an universal score, which corresponds to the degree of collocability. In this work the score is called *proximity* and ranges from 0 (absolutely independent terms) to 1 (perfect collocations). To get the proximity

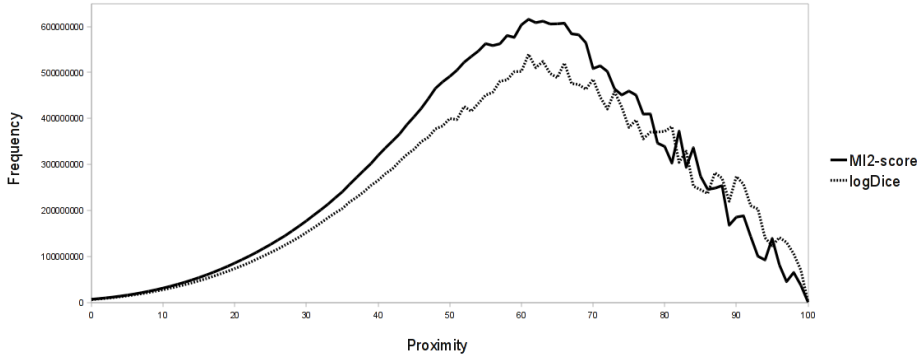

Fig. 1. Proximity distribution

from association scores described above, we need to transform the scores into interval $[0, 1]$. The conversion is done by normalizing the scores by their maximal values. The proximity distribution in a Czech web corpus (200 billions of tokens) for two best-resulting association scores is shown in Figure 1.

3 Domain Collocations

The association scores described in the previous section have some advantages – computation of their values is very fast and simple, they satisfactorily reflect collocation scores of global collocations, etc., but they also have at least one disadvantage – they are not convenient for identifying domain-specific collocations, which are relevant only for a small set of documents related to the same topic. An example of the domain collocation is *rozhodovací strom (decision tree)*, which is a strong collocation in the computer science domain but hardly in general.

The idea behind the method of identifying domain collocations is that domain collocations should be generated from domain specific sub-corpora. Nevertheless, there are many problems: what domains should be used, how to divide the corpus into domain specific sub-corpora, how to detect domain specific collocation, etc.

Probably the best way how to solve these problems is to avoid them. The solution is based on using the association scores described above with redefinition of the $f(t)$ function. Value of the $f(t)$ function is in the domain approach defined as *number of occurrences of term t in all documents containing all constituents of the investigated potential collocation*. In other words, for bigram (t_1, t_2), value of the $f(t_1)$ function is computed as a number of occurrences of t_1 in all documents containing both t_1 and t_2 at arbitrary location and vice versa. This approach has following consequences:

- Value of the $f(t)$ function is different for different bigrams. This is the source of a high computational complexity.
- Value of the domain proximity is always higher then value of the non-domain proximity for the same bigrams.
- Comparing to the non-domain approach, in the domain approach the proximity of good collocations increases rapidly, whereas proximity of non-collocations increases slightly.

Examples of proximity values for some bigrams from the corpus are shown in the Figure 2.

collocation type	**bigram**	**non-domain proximity**	**domain proximity**
global collocation	jízdní řády (timetables)	0.952	0.992
	karlovy vary	0.983	0.995
non-collocation	a ale (and but)	0.295	0.319
	zelené myšlenky (green ideas)	0.278	0.286
domain collocation	rozhodovací strom (decision tree)	0.363	0.684
	třecí síla (frictional force)	0.441	0.820

Fig. 2. Examples of proximity values.

4 Conclusions

One of the disadvantages of common association scores is a fact that some domain specific collocation cannot be identified. This work solves this problem by providing a new approach to computing term frequencies with regard to domain collocations. This method does not nearly affect scores of good general collocation and non-collocations but significantly improve proximity score in domain specific collocation. The proposed method has been tested and is being used in a real information retrieval system.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009.

References

1. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine in Practical Lexicography: A Reader. (2008) 297–306.
2. Yarowsky, D.: Word Sense Disambiguation. In: The Handbook of Natural Language Processing, New York: Marcel Dekker (2000).
3. Smadja, F., Hatzivassiloglou, V., McKeown, K.R.: Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics (1996).
4. Lin, D.: Using Collocation Statistics in Information Extraction. In: Proceedings of the Seventh Message Understanding Conference (MUC-7. (1998).
5. Church, K.W., Gale, W.A.: Concordances for Parallel Text. In: Proceedings of the 7th Annual Conference of the UW Center for the New OED and Text Research, Oxford, UK. (1991).
6. Oakes, M.P.: Statistics for Corpus Linguistics. Edinburgh University Press, Edinburgh (1998).
7. Dias, G., Guilloré, S., Lopes, J.: Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora. In: Proceedings of Traitement Automatique des Langues Naturelles (TALN), Cargèse, France. (1999).
8. Rychlý, P.: A Lexicographer-Friendly Association Score. In: Recent Advances in Slavonic Natural Language Processing. (2008).