Aggregation methods for efficient collocation detection

Anca Dinu, Liviu P. Dinu, Ionut T. Sorodoc

University of Bucharest,
Faculty of Foreign Languages and Literatures
Faculty of Mathematics and Informatics,
Centre for Computational Linguistics
14 Academiei, 010014, Bucharest, Romania,
anca_ d_ dinu@yahoo.com, ldinu@funinf.cs.unibuc.ro, ionut.sorodoc@gmail.com

Abstract

In this article we propose a rank aggregation method for the task of collocations detection. It consists of applying some well-known methods (e.g. Dice method, chi-square test, z-test and likelihood ratio) and then aggregating the resulting collocations rankings by rank distance and Borda score. These two aggregation methods are especially well suited for the task, since the results of each individual method naturally forms a ranking of collocations. Combination methods are known to usually improve the results, and indeed, the proposed aggregation method performs better then each individual method taken in isolation.

Keywords: collocation detection, language similarity, rank aggregation

1. Introduction

A collocation is defined as a sequence of two or more words that form a semantic unit and the whole has an independent existence beyond the individual parts.

In the words of Firth (1957): "Collocations of a given word are statements of the habitual or customary places of that word."

The importance of collocation detection is raised in one of the most recent study on collocations (Seretan, 2011). Collocations have a wide range of applications in natural language processing (such as in natural language generation, machine translation, parsing, word sense disambiguation, text classification or text summarization), thus extracting collocations was the subject of an impressive number of papers. An overview of such methods is given in (Manning and Schuetze, 1999) and (Pearce, 2002). Some good examples of popular collocation extraction systems which rely on statistic-based methods are Xtract (Smadja, 1993) and Colex (Orliac and Dillinger, 2003).

We propose in this paper a rank aggregation method for the task of collocations detection, which consists in aggregating by rank distance and Borda score the collocations rankings obtained by several traditional methods. Rank distance and Borda score are especially well suited for the task, as the results of each individual method naturally form a ranking of collocations. Since, in general, the multicriterial combination is superior to individual classification, combining collocation measures was investigated in previous work, such as in (Pecina, 2010).

We chose to employ (and re-implement) four well known individual collocation methods, on the basis of their good performance reported in the literature: likelihood ratio, Dice method, Z-test and χ-squared test.

The rest of the paper is structured as it follows. We present the methodology, then two experiments: in the first one, we analyse the performance of the aggregation methods on pos-tagged corpus and in the second one, we use these methods to analyse the language similarity. Finally, we draw the conclusions and suggest further investigation directions.

2. Methodology

Our approach is the next one :

- In the first part of our experiment, we verify the performance of the aggregation methods with a pos-tagged corpus and we follow these three steps:
  - we obtain the collocation rankings which are to be combined in a single ranking, using the four individual methods;
  - we define and compute a distance between pairs of rankings;
  - we determine the ranking which minimizes the sum of the distances from itself to all the initial rankings.
- After we verify the performance of the aggregation methods, we choose the best one: the rank aggregation method, to calculate the similarity between four languages using the Europarl corpus. We follow these steps:
  - we calculate a ranking of collocations using the rank aggregation method for every language;
  - we translate every collocation from rankings word by word from a source language to a target language;
  - we calculate the number of collocations that are the same in the initial ranking of a language and in the translated ranking from another language. The similarity is given by this number divided by the dimension of the rankings;

The collocation rankings are obtained by ordering the collocations by their score given by each particular method.

These scores can be interpreted as a confidence score for actually being collocations. Thus, the higher a collocation is positioned in the ranking, the more reliable is its collocation labelling.
Obviously, the combining individual methods may vary, both in number and in quality. We leave for further research experimenting with these choices.
In the remaining of this section, we briefly describe these four methods and explain the two aggregation method we propose: Rank distance aggregation and Borda score.

2.1. Contingency tables

An important notion used in the association measures is the contingency table for the observed data (see Table 1). This type of table contains four cells, which represent the frequency of the bigrams formed by $word_1$ and $word_2$ (O11), the number of bigrams with $word_1$ and without $word_2$ (O12), the number of bigrams with $word_2$ and without $word_1$ (O21) and the number of the bigrams without $word_1$ and $word_2$. On the basis of this table, one computes the contingency table for the expected data (see Table 2), which introduces the marginal frequencies, not only the observed frequency *O* and expected frequency *E*. Marginal frequencies are the sums of lines (R1 and R2) and columns (C1 and C2) from Table 1 and have an important role in statistical analysis and in the definition of the individual measures we use in this paper.

	$word2$	$\neg word2$
$word1$	O11	O12
$\neg word1$	O21	O22

Table 1: Contingency table for the observed data

	$word2$	$\neg word2$
$word1$	$E11 = \frac{R1*C1}{N}$	$E12 = \frac{R1*C2}{N}$
$\neg word1$	$E21 = \frac{R2*C1}{N}$	$E22 = \frac{R2*C2}{N}$

Table 2: Contingency table for the expected data

2.2. The individual measures

We give here the definitions of individual measures which we employe further in the paper.

1. Dice method: It was introduced by Smadja (Smadja, 1993):
$$DICE(w1w2) = \frac{2 * O_{11}}{R_1 + C_1}$$
2. Z-Test: A frequently used measure for collocation detection is the z-score:
$$Z - score(w1w2) = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$
3. χ^2 Test: It is an alternative statistical test which is not assuming that the probabilities are normally distributed:
$$\chi^2(w1w2) = \sum_{\substack{0<i<3 \\ 0<j<3}} \frac{(O_{ij} - E_{ij})^2}{N}$$
4. Likelihood ratio: It is based on the ratio between the likelihood values of observed data and the likelihood values of the expected data. The formula used in this paper based on the likelihood test is:
$$Likelihood(w1w2) = 2 * \sum_{\substack{0<i<3 \\ 0<j<3}} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

2.3. Rank Distance and Aggregation

A ranking is an ordered list of objects. Every ranking can be viewed as being produced by applying an ordering criterion to a given set of objects. We use a distance between two rankings to find a ranking which minimizes the distance to initial rankings.

Definition 1. *Rank distance is used to measure the similarity between two ranked lists (rankings). A ranking of a set of n objects can be represented as a permutation of the integers 1, 2, ..., n. S is a set of ranking results. $\sigma \in S$. $\sigma(i)$ represents the rank of object i in the ranking result . The rank distance is computed as (Dinu, 2002):*

$$RD(\sigma, \tau) = \sum_{i=1}^{n} |\sigma(i) - \tau(i)|$$

The ranks of elements are given from bottom up, i.e. from n to 1, in a Borda order. The elements which are not in one ranking receive the rank 0.
In a selection process rankings are issued for a common decision problem, therefore a ranking that "combines" all the original (base) rankings is required. One common sense solution is finding a ranking that is as close as possible to all the particular rankings. Apart from many paradoxes of different aggregation methods, this problem is NP-hard for most non-trivial distances.
Formally, the result of all the individually considered selection criteria is a finite collection of, not necessarily different, (partial) rankings, that we will call a ranking multiset $\mathcal{T} = \{\tau_1, \tau_2, ..., \tau_k\}$. When aggregating $\mathcal{T}$ into a single ranking we are looking for a σ with a minimal rank distance to all the rankings of the multiset; since Δ takes only positive values, we have to minimize the sum:

$$\Delta(\sigma, \mathcal{T}) = \sum_{\tau \in \mathcal{T}} \Delta(\sigma, \tau).$$

Definition 2. *Let $\mathcal{T} = \{\tau_1, \tau_2, ..., \tau_k\}$ be a multiset of rankings over object universe $\mathcal{U}$. A rank-distance aggregation (RDA) of $\mathcal{T}$ is a ranking σ (over the same universe $\mathcal{U}$) that minimizes $\Delta(\sigma, \mathcal{T})$. We denote the set of RD aggregations by $agr(\mathcal{T})$.*

A polynomial algorithm for rank aggregation is given by Dinu and Manea (Dinu and Manea, 2006). Rank aggregation was used with great results in other domains of computational linguistics, such as text categorization (Dinu and Rusu, 2010).

2.4. Aggregation with Borda score

The Borda method is similar to Rank distance, both of them being positional methods: they give a score to an element

Likelihood	Dice	z-score	χ - square	rank aggregation
i. e.	endoplasmic reticulum	endoplasmic reticulum	endoplasmic reticulum	alma mater
e. g.	cystic fibrosis	cystic fibrosis	cystic fibrosis	cystic fibrosis
same time	deja vu	deja vu	deja vu	deja vu
large number	myocardial infarction	myocardial infarction	myocardial infarction	myocardial infarction
twentieth century	conscientious objector	conscientious objector	conscientious objector	conscientious objector
other hand	adenylate cyclase	adenylate cyclase	adenylate cyclase	adenylate cyclase
black hole	thirteen colonies	thirteen colonies	thirteen colonies	thirteen colonies
recent year	angular momentum	coronary artery	coronary artery	i. e.
civil war	coronary artery	angular momentum	angular momentum	e. g.
early century	axial tilt	carbonic anhydrase	carbonic anhydrase	coronary artery

Table 3: Table with the 10-best collocations for every method described in section 2 and the results for rank aggregation.

according to the position in rankings. The main advantage of these methods is the simplicity of the implementation.
We consider an universe of elements U and a set of rankings $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$. This method take every element $c \in U$ and for every τ_i the Borda score is $B_i(c) =$ the number of elements which are beyond the element c in the ranking τ_i. The final Borda score for an element c is: $B(c) = \sum_{i=1}^{k} B_i(c)$. The final ranking is formed by sorting descending the elements of U by the value of final Borda score.

3. Experiments

This section is dedicated to the two experiments. The first one is designed to test the stability of our aggregation method for collocation detection on a pos-tagged corpus, namely the Wackypedia corpus (Baroni et al., 2009). The results confirm that aggregation methods we propose are competitive: they perform better then each individual method taken in isolation.
The second experiment is meant to investigate the similarities between languages, with respect to collocations. For this purpose we use the Europarl corpus (Koehn, 2005). We choose to analyse English, French, Italian and Spanish languages.

3.1. Aggregation on Pos-Tagged Corpus

The first experiment uses the Wackypedia corpus (about 100.000.000 words) (Baroni et al., 2009). We compute the frequency of each of these words and for each bigram. Using the preprocessed frequencies, we then calculate every individual collocation measure described in the Methodology section. In Table 3, it can be seen the top 10 rankings for the individual measures, as well as the rankings obtained after aggregation.
Some of the measures for collocation detection focus on cases of strong association, so we count only the collocation with $frequency \geq 10$. We limited our study only to bigrams formed with nouns and adjectives. We sort the lists for every method and we select the N-best collocations for every measure, thus obtaining a ranking of collocation for each individual measure. Finally, we aggregate these rankings by Rank aggregation and Borda score. These steps are depicted in Figure 1.
To validate the data we use the following method: for every ranking, we take every possible collocation and we verify its existence in WordNet. Taking *E* to be the set of the extracted collocations with a method, and *G* the set of collocations from WordNet, we calculate the precision of the method with the formula:

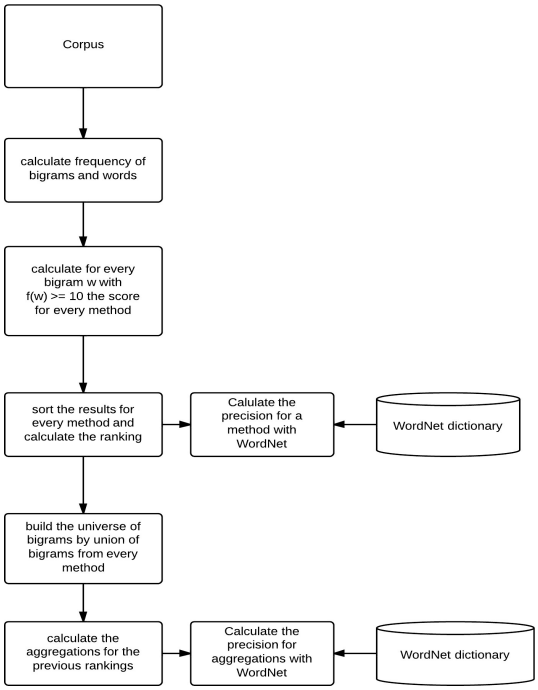

Figure 1: The steps followed by our algorithm

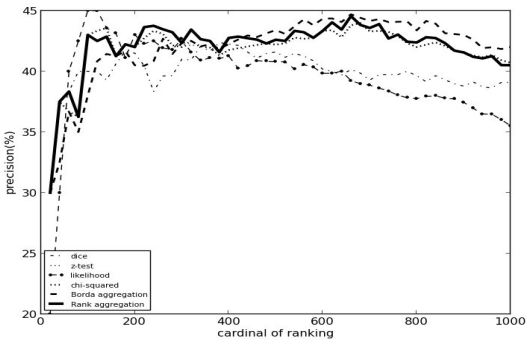

Figure 2: Graph with the evolution of the initial methods compared with aggregation methods

Method	Arithmetic Mean
Dice	40.38
Z Test	42.03
Likelihood	42.03
χ^2 Test	40.49
Borda aggregagtion	41.82
Rank aggregation	42.32

Table 4: Arithmetic mean of the precision on Wackypedia corpus

	1000	2000	3000	4000	5000
$Dice$	39.10	36.00	33.56	30.47	27.80
$ZTest$	40.70	36.80	33.43	31.15	29.24
$Likelihood$	40.70	36.80	33.40	31.50	29.08
χ^2	35.50	31.45	28.96	26.25	23.98
$Borda$	42.00	37.55	34.73	31.95	30.18
$Rank$	40.50	37.20	34.20	32.15	29.66

Table 5: Precisions for different cardinalities of rankings on Wackypedia corpus

$$p = \frac{|E \cap G|}{|E|} * 100$$

In figure 2 we observe the evolution for the four measures and for the two aggregation methods, function of the length of the rankings. One notices that both aggregation methods are constantly on top of the individual methods, with comparatively good precision.
The overall arithmetic mean of all measures, for values of the length of rankings between 20 and 750, is presented in table 4. The arithmetic mean is a standard tool to analyse the collocation detection. As one can see, the rank aggregation achieves good results and a better precision than all the individual measures. Also, Borda aggregation has a good position compared with the other measures.
In table 5, there are presented precisions for bigger values of the length of rankings. It can be observed that the aggregation methods have the best precisions in this case too.

3.2. Collocation detection and Language similarity

Since collocations are an important part of a language, we consider that a large number of collocations that can be translated word by word from a language to another is a good indicator that these languages are related. Other study on multilingual applications of collocation detection is developed in (Daudaravicius, 2010), but it has a different direction.
We chose to experiment with four languages: English, French, Italian and Spanish, all of which are part of Europarl corpus.
We create by aggregation methods presented in previous sections, rankings with 4000 collocations for every language. We translate word by word every collocation from the source language to the target language. We intersect the translated list of collocations with the collocation list generated by rank aggregation. This gives the number of overlapping (common) collocations of the two languages.

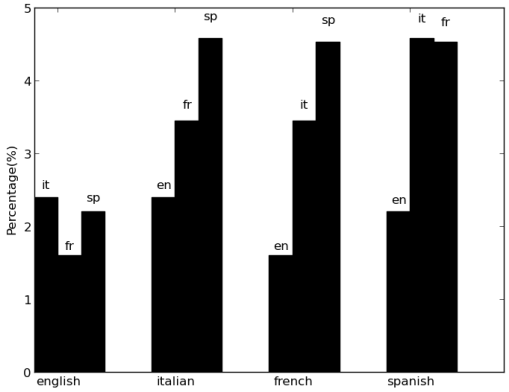

Figure 3: Collocations that can be translated from a language to another

The results are shown in Figure 3. As expected, English - Latin languages pairs have lower percentages than any other pair of romance languages. This approach can have applications in automatic translation or faster learning of a new language.
Here are some examples of collocations that are similar in different languages:

1. chantiers naval (fr) - cantieri navali (it)
2. highly skilled (en)- altamente qualificati (it)
3. compte tenu (fr) - tenendo conto (it)
4. disco compacto (sp) - disque compact (fr)
5. silencio ensordecedor (sp) - silence assourdissant (fr) - silenzio assordante(it)

We made publicly available the files containing the 4000 collocations for each language and the common collocations in each pair of languages.

4. Conclusions and future work

Results show that the aggregation methods are constantly better than the individual methods. Also, when used on four different languages, the aggregation method for detecting collocations was found to be consistent with the known language family relations (such as between romance languages).
One observes that it was nothing specific about the choice of initial set of individual methods. Hence, one might expect that being flexible about the initial set of individual method, can lead to even better precisions for collocation detection.
A notable advantage of the aggregation methods is that they can provide important information about association of individual measures. For instance, when the aggregation method obtains particularly good precisions on a set of individual methods, one can infer that the initial measures are complementary, mutually eliminating the weak points of each other.

We leave for further work improving the validation scheme in two respects. On the one hand, one might replace the WordNet collocation search, which is not very reliable, with some hand-made lists of human judgements on collocations. On the other, one might consider weighting the rankings such that finding in Wordnet a collocation placed at the top of the ranking (e.g. validating it as a collocation) should count more than validating a collocation placed at the bottom of the ranking. Also, it would be interesting to test these methods on different corpora, to see in what ways this choice influences the results.

Finally, we plan to extend our experiments to more languages, in order to investigate the behaviour of collocations in other family of languages (such as germanic, slavonic or hellenic).

5. Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. The contribution of the authors to this paper is equal. Research supported by a grant of the Romanian National Authority for Scientic Research, CNCS UEFISCDI, project number PN-II-ID-PCE-2011-3-0959.

6. References

M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43(3): 209-226, 2009.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):2229, March, 1990.

Vidas Daudaravicius. The Influence of Collocation Segmentation and Top 10 Items to Keyword Assignment Performance., in Alexander F. Gelbukh, ed., 'CICLing' , Springer, , pp. 648-660, 2010 .

Liviu P. Dinu. On the classification and aggregation of hierarchies with difererent constitutive elements. Fundamenta Informaticae, vol. 55, no. 1, 39-50, 2002.

Liviu P. Dinu and Florin Manea. An eficient approach for the rank aggregation problem. Theoretical Computer Science, vol. 359, no. 1, 455-461, 2006.

Liviu P. Dinu and Andrei A. Rusu: Rank Distance Aggregation as a Fixed Classifier Combining Rule for Text Categorization. CICLing, 2010 638-647.

Ted E Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 6174, 1993.

Stefan Evert(2004). The statistics of word cooccurrences: Word pairs and collocations. PhD Thesis, University of Stuttgart, 2004.

Stefan Evert and Brigitte Krenn. Methods for the qualitative evaluation of lexical association measures.Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, 2001.

John Rupert Firth. A synopsis of linguistic theory, 1930-1955, 1957.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. MT summit. Vol. 5. 2005.

Christopher D. Manning and Hinrich Schuetze. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, 1999.

Brigitte Orliac and Mike Dillinger. Collocation extraction for machine translation. Proceedings of Machine Translation Summit IX, 2003.

Darren Pearce (2002). A comparative evaluation of collocation extraction techniques. In Third International Conference on Language Resources and Evaluation (LREC), pages 15301536, Las Palmas, Spain, 2002.

Pavel Pecina. Lexical association measures and collocation extraction. Language Resources and Evaluation, Volume 44, Issue 1-2 , pp 137-158, 2010.

Violeta Seretan. Syntax-Based Collocation Extraction, volume 44 of Text, Speech and Language Technology. Springer, 2011.

Frank Smadja (1993). Retrieving collocations from text: Xtract. Computational Linguistics, 19(1), 143177, 1993.