PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **16** (2004), 171-182

PLISKA
STUDIA MATHEMATICA
BULGARICA

A STOCHASTIC APPROACH FOR FINDING OF SEMANTICALLY RELATED WORDS

Veska Noncheva, Pablo Gamallo, Alexandre Agustini, Gabriel Lopes

Semantically related words are modelled as words having the same probability distribution on the set of syntactic contexts occurring in text corpora. A learning algorithm for finding of clusters of semantically related words is developed. In that algorithm χ^2 statistics is used as a performance measure.

1. Introduction

This paper describes some results of our efforts to develop a learning algorithm for automatic acquisition of clusters of semantically related words from text corpora.

Each word in the corpus possesses a certain quantity of context.

Context could be revealed through a handful of syntactic features. Some of them are:

- selection restrictions – the semantic constraints that a word needs to match in order to be syntactically dependent and attached to another words (see [2]);
- contiguous and non-contiguous multiword lexical units (see [8]);
- verb subcategorizations (see [7]).

2000 *Mathematics Subject Classification:* 62P99, 68T50
Key words: syntactic context, semantic preferences, χ^2 goodness of fit test

Context could be found by combining the results of several restricted and complementary processes.

We have found for each word in the corpus a certain number of grammatical clues. We consider these grammatical clues as features of words. These features form the recognized *syntactic context* of a word.

We could also use these clues for each word to determine if two or more words are used in a lexically similar manner. Words that are used in a similar way throughout a corpus are indeed semantically related.

Our aim is to discover both classes with semantically related words and clusters of syntactic contexts that share the same selection restrictions.

Our approach is partially based on Harris' Distributional Hypothesis, which states that words that occurred in the same contexts tend to have similar meanings [4]. In our paper we raise the stronger hypothesis that words having the same stochastic behaviour on the syntactic contexts throughout a corpus are indeed semantically related. In addition we raise the hypothesis that syntactic contexts having the same stochastic behaviour on the words throughout a corpus are indeed semantically related and they share the same selection restrictions.

2. Syntactic Contexts and Semantic Information

In order to define syntactic context at first we will define the notion of syntactic dependency between words.

Let w_1 and w_2 be two words. A *syntactic dependency* between w_1 and w_2 is the triplet $< r, w_1, w_2 >$, where r represents a grammatical relation between these two words. For example: < dobj, ratify, treaty >, and < of, republic, Portugal > are the dependencies extracted from the expressions *to ratify the treaty* and *republic of Portugal.* The grammatical relation$r$can be a direct object (dobj), a subject (subj), a modifier, a preposition, etc. Some of these relations are identified by a set of syntactic features such as the relative position between two words. For instance, observing a noun phrase at the left of a verb phrase may identify the subject relation in English.

A *syntactic context* is extracted by abstracting one of the related words of the syntactic dependency. Thus, two complementary syntactic contexts can be abstracted from a syntactic dependency (*see* [1,2]). For example: both **republic-of** and **of-Portugal** are the two contexts abstracted from the second dependency introduced above.

The main property of a syntactic context is that it imposes *selection restrictions* (i.e., *semantic preferences*) on the words with which it combines. The *set of all words* that fill the semantic conditions of a context represents the extensional definition of the *selection restrictions* imposed by the context.

Our aim is to discover clusters of contexts imposing the same *semantic preferences* and classes with words found as being semantically related. For example:

- Cluster of contexts requiring legal documents:**ratification-of**, **dobj-ratify**, **dobj-approve**, **dobj-sign**, **signatories-to**, and **legislation-of**. Words appearing in these contexts and found as being semantically related: *article*, *law*, *decree*, *document*, and *treaty*.

- Cluster of contexts requiring countries: **north-of**, **factory-in**, **made-in**, and **republic-of**. Words appearing in these contexts and found as being semantically related: *Ireland*, *Portugal*, and *France*.

- Cluster of contexts requiring activities: **in-course**, **interruption-of**, and **end-of**. Words appearing in these contexts and found as being semantically related: *work*, *procedure*, and *travel*.

3. Model

Let $\{1, 2, \ldots J\}$ be a finite set of syntactic contexts. Let W be a categorical random variable defined on the set of syntactic contexts. Its levels $\{1, 2, \ldots J\}$ do not have a natural ordering so W is a nominal random variable. We will call it *polysemantic word* or simply *word*.

Let $\{1, 2, \ldots I\}$ be a finite set of natural language words. Let S be a nominal random variable defined on this set of words. We will call S *context*.

Let W_1 and W_2 be two words defined on the set $\{1, 2, \ldots J\}$. We will say that W_1 and W_2 *are used in a lexically similar manner* or simply that W_1 and W_2 *are related*, if they have the same probability distribution on the set of syntactic contexts.

Let S_1 and S_2 be two contexts defined on the set of natural language words $\{1, 2, \ldots, I\}$. We will say that S_1 and S_2 are *similar*, if they have the same probability distribution on the set of words.

Let W and S be two nominal random variables called word and context, respectively. A bivariate relationship between them is defined by their *joint probability distribution* $\{\pi_{ij}\}$, where π_{ij} denotes the probability that (W, S) falls in the cell in row i and column j of the rectangular table that displays this distribution. The interpretation of π_{ij} is that it is the probability that the word i has the syntax j.

When W is fixed rather than random, the notion of a joint distribution for W and S is no longer meaningful. However, for a fixed level of W, the random variable S has a probability distribution. It is appropriate to study how the probability distribution of the context changes as the word changes.

Let $\{\pi_{ij}\}$ denote the probability of classification of the given word i as a word having syntactic context j, $j = 1,\ldots, J$. Then $\sum_j \pi_{j/i} = 1$. The probabilities $\{\pi_{1/i}, \pi_{2/i}, \ldots, \pi_{J/i}\}$ form the *conditional distribution* of the syntax S for the given word i.

We will compare the conditional distribution of the context S at various words from W.

When a random sample of n appearances of a word is taken from a probability distribution concentrated on a set of J syntactic contexts the number of words $\{n_j\}$ in each syntactic context have multinomial $(n, \{\pi_j\})$ distribution, characterized by the sample size n and the cell probabilities $\{\pi_j\}$, $j = 1, \ldots, J$.

Pearson's goodness-of-fit test evaluates whether probabilities in a multinomial distribution equal certain values.

Consider the null hypothesis H_0 that J parameters of the multinomial distribution equal certain values $\{\pi_j^0\}$, where $\sum_j \pi_j^0 = \sum_j \pi_j = 1$.

When the χ^2 statistic is too small, the fit seems too good.

We will apply Pearson's goodness-of-fit test to find classes with related words and clusters with similar contexts.

Let $(n^h, \{\pi_j^h\})$ be a fixed multinomial distribution. We will use Pearson's χ^2 goodness-of-fit test in order to compare the distributions of the word i, $i = 1, \ldots, I$, with this $(n^h, \{\pi_j^h\})$ multinomial distribution. In this way we will find the set of words having the feature: The counts of the words in each syntactic context have the same $(n^h, \{\pi_j^h\})$ multinomial distribution. Let the distribution $(n^h, \{\pi_j^h\})$ run the given finite set of multinomial distributions $\left\{(n^h, \{\pi_j^h\})\right\}_{h=1}^{I}$. Then we will obtain different classes with related words.

Let $(n^k, \{\pi_i^k\})$ be a fixed multinomial distribution. We will use Pearson's χ^2 goodness-of-fit test in order to compare the distributions of the context $j, j = 1, \ldots, J$, with this $(n^k, \{\pi_i^k\})$ multinomial distribution. In this way we will find the set of contexts having the feature: The counts of the contexts of each word throughout the corpus have the same $(n^k, \{\pi_i^k\})$ multinomial distribution. Let now the distribution $(n^k, \{\pi_i^k\})$ run the given finite set of multinomial distributions $\{(n^k, \{\pi_i^k\})\}_{k=1}^{J}$. In this way we will obtain clusters with similar contexts.

As additional results we will know both the probability distribution of every one word and recognized context from the corpus.

4. Algorithm

We are proposing an algorithm finding sets of words having the same population distribution and the set of their probability distributions.

Input: sample multinomial distributions, level of significance.

Output: set L of sets L_i of words having the same population distribution and the set of their probability distributions.

Initialisation: $i = 1$.

1. $L_i\{\text{word } i\}$, $k = i$. Go to 2.
2. Generate a multinomial distribution $(n, \{\pi_j^o\})$ using the sample distribution of word i. Go to 3.
3. Test the goodness of fit hypothesis $\mathrm{H}_0 : \pi_{j/i} = \pi_j^o$, $j = 1, \ldots, J$. If H_0 cannot be rejected go to 4 else go to 2 in order to generate a new multinomial distribution.
4. $k = k + 1$. If $k \leq I$ go to 5 else go to 6.
5. Compare the distribution $(n, \{\pi_j^o\})$ with distribution $(n, \{\pi_{j/k}\})$, by testing goodness-of-fit hypothesis:
   $\mathrm{H}_0^k : \pi_{j/k} = \pi_j^o$, $j = 1, \ldots, J$.
   If we cannot reject the hypothesis then $L_i = L_i \cup \{\text{word } k\}$. Go to 6. We used χ^2 test to find the set of words having multinomial distribution, presented by the probabilities $\{\pi_j^o\}$, $j = 1, \ldots, J$.
6. $i = i + 1$. If $i \leq I$go to 1, else go to 7.
7. End.

Let us point some features of the algorithm proposed above:

- Algorithm developed for computers. It draws on concepts and results from statistics.
- Learning by experience algorithm. One of the algorithm's tasks is an appropriate probability distribution to be generated. For this purpose probability estimations are calculated. When there are empty cells extremely small constants have been added. The obtained solution is evaluated by a goodness-of-fit test and after that the generation process could be repeated under new information obtained. Therefore this algorithm automatically improves its behaviour and this improvement is based on the evaluation of its heuristic solutions.

5. Reduction of the number of the syntactic contexts

We will treat cells counts n_{ij} of the two-way table as random variables with expected values denoted by $m_{ij} = \mathrm{E}n_{ij}$. Let S be considered as the explanatory variable and W be considered as the response. We will study how the probability distribution of the word W changes as the context S changes. Let totals n_{+j} be fixed. We regard each column j of I counts n_{ij}, $i = 1, \ldots, I$, as an independent multinomial sample on W. Then the cell counts $\{n_{ij}, i = 1, \ldots, I\}$ have multinomial distribution with response probabilities $\{\pi_{i/j} = m_{ij}/m_{+j},\ i = 1, \ldots, I\}$, where m_{ij} are expected frequencies. Cell counts from different columns are independent. By analogy applying the algorithm, described above, we will obtain classes of similar contexts and will reduce the number of syntactic contexts.

6. How do we mine selection restrictions?

Linguistic hypothesis: The syntactic contexts considered as similar impose the same selection restrictions. Selection restrictions could be revealed through a list of words that may appear in similar contexts. This linguistic hypothesis is discussed in [2].

We consider that the words that have the same probability distribution $\{\pi_{j/i}\}_{j=1}^{J}$ on the set of syntactic contexts belong to the same semantic class L_i. Both the semantic class L_i and the probability distribution $\{\pi_{j/i}\}_{j=1}^{J}$ on the set of similar semantic contexts represent the selection restrictions of those contexts.

7. Example

The data about the words and recognized syntactic contexts obtained from an artificial corpus are shown in Table 1. The results from calculation $\chi 2$ statistics for data given in Table 1 are given in Table 2. Interpreting these results with chosen level of significance .05 we receive two clusters with similar contexts that are presented in Table 3.

The results from calculation $\chi 2$ statistics for data given in Table 1 in order to obtain classes of related words are shown in Table 4. Interpreting these results when the level of significance is .05 we receive three classes with related words that are presented in Table 5.

Let us summarise the results. Applying the algorithm we receive:

Cluster of syntactic contexts: **dobj-cook, dobj-eat**. The classes of words appearing in these contexts are: $\{$*eggs, fish, sausage*$\}$ and $\{$*things*$\}$.

Cluster of syntactic contexts: **subj-cook**, **subj-eat, subj-write**. The classes of words appearing in these contexts are: $\{$*Maria, Paulo, Pedro*$\}$ and $\{$*Marta*$\}$.

	dobj-cook	dobj-eat	dobj-lookfor	dobj-read	dobj-write	subj-cook	subj-eat	subj-lookfor	subj-read	subj-write	
Maria			3			26	23	15	40	17	124
Marta						2	3	1		3	9
Paulo			1			3	5	7	15	4	35
Pedro			8			16	31	35	42	27	159
book			5	35	7						47
eggs	50	30	5								85
fish	6	4	1								11
letter			1	7	21						29
novel			2	9	18						29
sausage	43	53	8								104
things	4	7	9	5	4			1			30
	103	94	43	56	50	47	62	59	97	51	662

Table 1: Cell frequencies

Class of semantically related words {*Maria, Paulo, Pedro*}. The clusters of contexts appearing with these words are: [**dobj-lookfor**], [**subj-cook**, **subj-eat, subj-write**], [**subj-lookfor**], and [**subj-read**].

Class of semantically related words {*eggs, fish, sausage*}. The clusters of contexts appearing with these words are: [**dobj-cook, dobj-eat**] and [**dobj-lookfor**].

Class of semantically related words {*letter, novel*}. The contexts appearing with these words are: **dobj-lookfor**, **dobj-read,** and **dobj-write**.

8. Related work on semantic acquisition

In Grefenstette and Lin's approach words that share a great number of features are found as being similar. Similarity measure between words can be calculated. This measure takes into account:

The numbers of features that two objects do or do not share.

The importance of these features for each word.

This approach is presented in [3] and [6].

Pablo Gamallo and Alexandre Agustini use a deterministic method for semantic information extraction merely based on word cooccurrence within basic syntactic constructions. Their strategy relies on two basic linguistic assumptions. First, they assume that two syntactically related words impose semantic selectional restrictions to each other. Second, it is also claimed that two syntactic contexts impose the same selection restrictions if they cooccur with the same

	1.dobj-cook	***2.dobj-eat***	3.dobj-lookfor	4.dobj-read	5.dobj-write
1.dobj-cook	0.01	**15.20**	218.01	2384274.42	2128837.73
2.dobj-eat	**13.66**	0.01	158.78	2219338.17	1981572.89
3.dobj- lookfor	249143.88	227353.07	0.00	213587.89	190709.39
4.dobj-read	2492357.01	2274583.66	202.99	0.01	111.96
5.dobj-write	1676912.20	1530388.26	478.55	66.05	0.01
6.subj-cook	2071051.19	1890093.74	3856.33	1126051.19	1005412.89
7.subj-eat	2531916.52	2310690.71	6403.03	1376626.19	1229142.32
8.subj-lookfor	2618763.47	2389949.70	871.89	1423847.97	1271305.62
9.subj-read	3811162.00	3478162.00	336.93	2072162.00	1850162.00
10.subj-write	2106543.14	1922484.31	7708.96	1145347.06	1022641.18

	6.subj-cook	**7.subj-eat**	8.subj-lookfor	9.subj--read	**10.subj-write**
1.dobj-cook	2008419.40	2649341.73	2511957.72	4144869.89	2179332.02
2.dobj-eat	1887187.04	2489421.09	2338206.44	3894674.00	2047782.79
3.dobj- lookfor	219692.51	289807.50	164739.59	453417.10	238388.81
4.dobj-read	1158330.79	1527973.64	1427704.81	2390498.29	1256902.21
5.dobj-write	780266.20	1029266.20	960586.56	1610282.80	846666.20
6.subj-cook	0.01	**6.80**	25.40	8257.70	**10.33**
7.subj-eat	**8.69**	0.01	9.66	14078.91	**0.49**
8.subj-lookfor	819.03	1054.43	0.01	3291.98	867.07
9.subj-read	22.60	12.61	18.10	0.01	16.43
10.subj-write	**10.30**	**0.41**	8.45	17116.85	0.01

Table 2: Result from calculation $\chi 2$ statistics for data given in Table 1 in order to obtain classes of similar contexts.

Class 1	direct_object-cook, direct_object-eat
Class 2	subject-cook, subject-eat, subject-write

Table 3: Classes of similar contexts.

	1**<u>Maria</u>**	2Marta	3**<u>Paulo</u>**	4**<u>Pedro</u>**	5book	6**eggs**
1**Maria**	0.00	1167837.09	31.58	**24.32**	12580144.69	22751274.27
2<u>Marta</u>	<u>6.72</u>	0.00	<u>12.49</u>	<u>6.94</u>	1201120.00	2172231.11
3**<u>Paulo</u>**	<u>6.09</u>	581163.87	0.00	**5.18**	4350887.21	7868601.71
4**Pedro**	**29.43**	1034735.44	**20.71**	0.00	14587797.90	26382140.58
5book	33611998.31	2487565.64	9487314.09	43099220.93	0.00	23040577.32
6**eggs**	49600087.16	3626627.35	14000085.29	63600080.85	18800197.76	0.00
7**fish**	5861829.8	433654.27	1654556.5	7516373.3	2221841.1	**0.23**
8***letter***	20951764.15	1523865.69	5913832.90	26865556.51	75.70	14362158.83
9***novel***	17317273.94	1269365.97	4887962.72	22205202.02	51.06	11870760.76
10**sausage**	55537792.90	4086514.58	15676058.23	71213741.23	21050792.23	**13.21**
11things	4381429.3	561007.70	1236745.4	5618037.9	1034015.4	1190031.2

	7**fish**	8letter	9novel	10**sausage**	11things
1**Maria**	2944338.35	7762242.89	7762241.83	27836838.50	7485514.36
2<u>Marta</u>	281120.00	741120.00	741120.00	2657786.67	733337.44
3**<u>Paulo</u>**	1018315.83	2684602.06	2684601.64	9627458.74	2357181.38
4**Pedro**	3414213.64	9001013.33	9001007.50	32279308.78	7000166.81
5book	2981850.72	77.85	46.38	28190787.96	118.98
6**eggs**	**1.28**	11600203.53	11600199.26	**10.74**	181.95
7**fish**	0.00	1370933.8	1370932.5	**0.95**	<u>20.08</u>
8***letter***	1858710.34	0.00	**1.45**	17572503.52	95.31
9***novel***	1536277.17	**2.00**	0.00	14524208.48	72.01
10**sausage**	**9.65**	12988881.21	12988872.29	0.00	147.15
11things	154014.97	638067.89	638028.10	1456019.40	0.00

Table 4: Result from calculation $\chi 2$ statistics for data given in Table 1 in order to obtain classes of related words.

Class 1	Maria, Paulo, Pedro
Class 2	eggs, fish, sausage
Class 3	letter, novel

Table 5: Classes of related words.

words. This approach is discussed in [2].

9. Related clustering approach

We will compare our approach with the two-way clustering approach [5].

The data structures in both approaches are standard. They assume cases and variables.

For clusters of both cases and variables, the basic unit is a block. A block is a sub- matrix of the data matrix. The data matrix could thus be reduced.

Sometimes cases are clustered and sometimes variables are clustered. This is a traditional classification scheme. In our approach this scheme explicitly connects clusters of words and clusters of contexts. It produces two types of clusters – clusters of words and clusters of contexts.

The two-way clustering scheme could simultaneously produce three types of clusters: the tree formed by the case clusters, the tree formed by the variable clusters and the data clusters formed by the blocks themselves. The data clusters could form a tree or a partition (see [5]).

The result of our algorithm is two clusters with the same clustering structure – partitions. A tree structure of any cluster could also be built. Here the search for closest pairs is rather expensive.

In both our and two-way clustering approaches the clusters are produced without a once-and-for-all distance calculation.

However the main difference between these two approaches is the following: Our approach is based on the probability model, described above. The two-way techniques themselves are not based on sound probability models.

Conclusion

Our approach is based on the assumption that both the words and the contexts in natural language processes are governed by probability distributions and that reasonable decisions can be made by reasoning about these probabilities together with observed data. This stochastic approach provides the basis for a learning algorithm. The results obtained are:

Classes with related words. As an additional result we obtain the probability distributions of the related words on the set of contexts. We could use these probabilities to govern some linguistic processes, such as the parsing process, for example. We could interpret a class with related words in the following manner:

Other words in the class give a clue to the meaning of a word from this class.

Other words in the class describe (add information to) the context of a word from this class.

Other words in the class present the context of one unknown word from this class.

Clusters with similar contexts. We are able to reduce the number of word contexts. As an additional result we obtain the probability distributions of the contexts on the set of words.

We apply the algorithms presented above in the following order: First, contexts are put in clusters. Each cluster is constituted by those contexts imposing the same selection restrictions. Second, words are put in classes. Each class contains those words filling the same selection restrictions.

We follow this order because contexts are less ambiguous than words. So, it is easier to build semantically homogenous clusters of contexts than to generate clusters of semantically homogenous words.

Our method for finding of semantically related words has two properties. The first concerns the extraction of corpus-specific semantics. When the corpus from which the information is derived defines a domain, the relations discovered are specific to that domain. The second one is that the results become more stable as more word's contexts are added for any one word.

REFERENCES

[1] P. Gamallo, C. Gasperin, A. Agustini, G.P. Lopes. Syntactic-Based Methods to Measure Word Similarity, Text, Speech, and Discourse (TSD'01), 2001, LNAI, Springer-Verlag, pp 116–125.

[2] P. Gamallo, A. Agustini, G.P. Lopes. Using Co-composition for Acquiring Syntactic and Semantic Subcategorisation, ACL-SIGLEX'02, 2002, Philadelphia, USA, pp 34–41.

[3] G. Grefenstette. Explorations in automatic thesaurus discovery, 1994, Kluwer Academic Publishers, USA, 305 p.

[4] Z. Harris. Distributional Structure, In J.J. Katz (ed.), The Philosophy of Linguistics, 1985, New York: Oxford University Press, pp 26–47.

[5] J. Hartigan. Clustering Algorithms, 1975, Wiley-Interscience, New York.

[6] D. Lin. Automatic Retrieval and Clustering of Similar Words, COLING-ACL'98, 1998, Montreal, Canada, pp 768–774.

[7] N. Marques, G.P. Lopes. Mining subcategorisation information by using multiple feature loglinear models, *10^{th} CLIN Meeting* (2000), 117–126.

[8] V. Noncheva, J.F. Silva, G. Lopes. Automatic acquisition of word interaction patterns from corpora. *10^{th} Conference of the European Chapter of the Association for Computation Linguistics EACL-03, Proceedings of the Workshop on Language Modeling for Text Entry Methods* (2003), 25–32.

Veska Noncheva
Dept. of Applied Mathematics and Modelling
University of Plovdiv, Bulgaria

Pablo Gamallo
Dept. of Computer Science
New University of Lisbon, Portugal

Alexandre Agustini
Dept. of Computer Science
New University of Lisbon, Portugal

Gabriel Lopes
Dept. of Computer Science
New University of Lisbon, Portugal