CREATIVE MATH.
13 (2004), 79 - 83

Towards Clustering from Corpora by Using Pattern Oriented Approaches

Dana Avram Lupşa

Abstract. Clustering from untagged corpora is very important especially for languages that have no such hierarchies as WordNet. The idea is to combine automatic noun clustering from unannotated corpora with some supervised learning methods. This paper presents a study on automatic noun clustering from texts selected from corpora by using a pattern-oriented filter. The patterns used are oriented to Romanian language but can be extended also to other languages. A comparison between results obtained by not applying any pattern and by applying different patterns (as filters) is also presented.

1. Introduction

The goal of extracting semantic information from text is well established, and has encouraged work on lexical acquisition (Roark and Charniak, 1998), information extraction (Cardie,1997), and ontology engineering (Hahn and Schnattinger, 1998). The purpose of this kind of work is to collect information about the meanings of lexical items or phrases, and the relationships between them, so that the process of building semantic resources (such as WordNet) by hand can be automated or at least helped.

Words clustering can be useful in construction of a set of synonyms for word sense disambiguation, to perform query expansion in QA systems (Oraşan & others, EACL), to build ontology from text [10], for data mining [11], especially for languages other than English, for which does not exist hierarchies such Wordnet, as in Romanian language case.

The main results to date in the field are concerned with extracting lists of words that belong together to a particular category(Riloff and Shepherd, 1997) (Roark and Charniak, 1998).

This paper describes three methods to discover semantic relation between words from unannotated corpus. The experiment is applied to Romanian language, but can be extended to other languages as well. Applied to a corpus of 200000 words, the methods reach a precision up to 70%.

This paper is structured as follows: in section 2 we present the technique of grouping words, that appears in the same context, into one cluster. Sections 3 and 4 we suggest some sets of clustering patterns to be applied to Romanian language. Some experiments on a 200000 Romanian corpus are discussed as well. The article finishes with a discussion about the results and some future research directions.

Received: 12.09.2004. In revised form: 27.11.2004.
2000 *Mathematics Subject Classification.* 68T50, 91C20.
Key words and phrases. *Natural language processing, semantic information, clustering.*

2. Same context

Most work on automatic lexical acquisition has been based at some point on the notion of semantic similarity. The underlying claim is that words that are semantically similar occur with similar distributions and in similar contexts (Miller and Charles, 1991).

According to this, the most intuitive method used in literature is to consider different nouns that appear in the same context window. By applying the method to our corpus, we get the results presented in the next table:

Context window (dimension)	**Words no. (in clusters)**	**No. of Clusters**	**Accuracy**	
			Clusters	**Multi-Clusters**
2+4	126	47	21%	38%
4+4	61	24	**54%**	54%
5+5	46	18	50%	55%

3. Prepositions

Each preposition adds some meaning when linking the features of the sentence. Solving the prepositional attachment problem usually is not a simple task because there are many possible directions the prepositions can be adjoined on (Whittemore, 1990). We believe this is the reason the prepositions (, as far as we know,) are not used to discover semantic relations among words.

In Romanian language, the preposition ***de*** (a sort of ***of***) has a special place because usually it has priority over other prepositions and it is attached to the word that precedes it. When it introduces the attribute of a noun, it comes to complete the meaning of that word. For example, the expression ***de brad*** (***of fir***) in the expression ***o ramura de brad*** (***a fir branch***), identifies the type of object we are referring to and that's why it play an important semantic role. The *de-expression*s express possession, species, material, author or cause. Some of the *de-expression*s (preposition ***de*** and the word that first comes after ***de***) are enough specialized so that they are attached only to a small set of semantically related words.

For example, on a corpus of 200000 words, the expression ***de albine*** (***of bees***) get two nouns:

miere, roi (honey, swarm)

but some less specialized de-expressions get a union of semantically related words, as in the next examples:

de-expression		*nouns*	
de acasa	***from home***	***treburile, lucruri, cumparaturi, reviste***	***business, things, shopping, magazines***
de aur	***of gold***	***cerbul, cornului, gura, medalia, bani, perioada***	***stag, antler, mouth, medal, money, period***

We expect that the expressions with high potential cannot appear with many nouns. But if a noun is not semantically related with the themes of the text, it can appear by accident.

So, one problem is when *de-expression* appears too frequent and with many nouns, and the other of it appears too seldom, for example 2 nouns are grouped together by only one de-expression. So, one the problem is to choose between two less and too many. The solution we suggest is to get only the *de-expression* whose apparition is under a certain rank, and then to calculate a similarity value based on co-occurrences of the *de-expression*.

The experiment

We applied the next two patterns

<*noun*> ***de*** <*word*> (*de-exp* **pattern 1**)

<*noun*> ***de*** <*noun*> (*de-exp* **pattern 2**)

on a corpus of 200000 words. The nouns that have as attribute the same *de-expression* considered being part of a group of clusters (multi-cluster). We manually evaluated the obtained groups. The precision of the clusters built in this way is given in the next table:

	Words no. in clusters	**No. of Clusters**	**Accuracy**	
			Clusters	**Multi-clusters**
de-exp **pattern 1**	157	55	25 %	54 %
de-exp **pattern 1 & 2 words cluster**	34	21	33 %	57 %
de-exp **pattern 2**	108	38	28 %	47 %
de-exp **pattern 2 & 2 words in cluster**	26	14	**42 %**	71 %

As we can see, the more refined pattern, improved with POS information, gets a precision of 42% as correct clusters, and 71% as being a multi-cluster.

On our corpus, we cannot compute similarities based on co-occurrences, because we get only one group of nouns that appears with more than one *de-expression.* This group is:

timp – perioada ***(time – period)*** **Experiment with larger size context**

The *de-expression* ***de albine*** identify the next set of nouns:

miere, stup, ceara, venin, intepatura, familii, colonia, specii, crescator, spor	***(honey, beehive, wax, venom, prick, families, colony, species, breeder, increase)***

Collecting the *de-expression* the word list gets and then, applying the de-expression methods, we get the next set of clusters:

C1: ***miere, stup, ceara***
C2: ***venin***
C3: ***intepatura***
C4: ***familii***
C5: ***colonia, specii, crescator***
C6: ***spor***

In our opinion, the result is good. Among those clusters, only the **C5** is erroneous. A ***breeder*** is someway linked to ***colonies*** and ***species***, by the nature of the activity, but a human subject would consider him to be semantically different and would put him in a separate group.

4. The power of enumeration

If a coordination conjunction appears between two words, they can indicate an enumeration of objects/concepts that are related. Possible enumerations are described in pattern 1.

Enumeration pattern 1 (weak condition)

<word> ([***, /si/sau***] *< word >*) $^{+}$

and if all separators are comma – there must be at least 3 words in the pattern

Comma is used as separator in other many contexts, so we introduce one more restriction in the pattern, in order to limit the enumeration identification errors.

We say that: if before the first word in the pattern, there is an enumeration separator, there is a higher probability for the word to be part of that enumeration. We supplement, in this way, the lack of a syntactic analyzer. This observation is illustrated by pattern 2:

Enumeration pattern 2 (semi-strong condition)

(the fist word will be ignored)

<word_to_be_ignored> [***, /si/sau***] *<word>* ([***, /si/sau***] *< word >*) $^{+}$

Our pattern identifies enumerations of things expressed by only one word. But sometimes happens that the last word in the pattern starts syntactic phrase formed by more than one word. In this case, only the head of the phrase is part of the enumeration. Usually, the first word of a phrase is not the head of that phrase. If we try to minimize the errors, we can ignore the last word, as illustrated in the pattern 3:

Enumeration pattern 3 (strong condition)

(the fist and last word will be ignored)

<word_to_be_ignored>[***, /si/sau***] *<word>* ([***, /si/sau***] *< word >*)$^{+}$
[***, /si/sau***] *<word_to_be_ignored>*

We apply those patterns on a corpus of 200000 words and then manually evaluated the clusters. The number of identified clusters and their precision are presented in the next table:

	No. of clusters	**No. of words**	**Accuracy**		**Observation**
			Multi-clusters	**Clusters**	
enumeration pattern 1 weak condition	1689	742	39 %	17 %	the evaluation was made on 100 clusters randomly chosen
Enumeration pattern 2 semi strong cond.	453	150	48 %	23 %	evaluation was made on all clusters
Enumeration pattern 3 strong condition	169	72	76 %	**70 %**	evaluation was made on all clusters

5. Conclusion and future research

Discussion of results

This paper presents some methods to select semantically related words from unannotated text and get precision comparable with other modern methods. Experimental evaluation suggests that the measure performs encouragingly well (an accuracy up to 70% for clusters). This is close to the best results reported in literature (accuracy of 78%) in condition that they use larger, and POS and syntactic annotated corpora (Resnik, 1995), (Cederberg and Widdows, 2003).

Future research

The methods get information about different set of words. For example, *de-expression* compared with enumeration method obtain only 26 common words. There are 131 words in *de-expression* that does not appear in *enumeration*, and there are 1663 in *enumeration* that does not appear in *de-expression.* Our intention is to combine those different methods into one, more powerful as precision and as the number of words that are grouped into clusters.

References

[1] Avram Lupşa D., Şerban G., Tătar D., *From Noun's Clustering to Taxonomies on an Untagged Corpus*, Babes-Bolyai University of Cluj-Napoca, Faculty of Mathematics and Computer Science, Research Seminars, Seminar on Computer Science, 182-192, 2003

[2] Cardie C., *Empirical Methods in Information Extraction*, AI Magazine, 18, 65–79, 1997

[3] Cederberg S., Widdows D., *Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction*, CoNLL, 2003

[4] Hahn U., Schnattinger K., *Towards Text Knowledge Engineering* , AAAI/IAAI, 524–531, 1998

[5] Miller, G., Charles, W., *Contextual correlates of semantic similarity* Language and Cognitive Processes, 1991

[6] Oraşan, C., Tătar, D., Şerban G., Avram, D., Oneţ A., *How to Build a QA System in Your Back-Garden: Application for Romanian*, EACL 2003

[7] Resnik, Ph., *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*, 1995

[8] Roark B., Charniak E., *Noun-phrase Co-occurrence Statistics for Semi-automatic Semantic Lexicon Construction*, COLING, 1998

[9] Whittemore, G., Ferrara, K., Brunner, H., *Empirical Study of Predictive Powers of Simple Attachment. Schemes for Post-modifier Prepositional Phrases*, 1990

[10] http://www.semantxls.com/ka.shtml

[11] http://www.research.ibm.com/compsci/kdd/

Babes Bolyai University
Computer Science Department
Kogalniceanu 1, 400084 Cluj-Napoca, Romania
E-mail address: davram@cs.ubbcluj.ro