STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **L**, Number 2, 2005

SOME REMARKS ABOUT FEATURE SELECTION IN WORD SENSE DISCRIMINATION FOR ROMANIAN LANGUAGE

DANA AVRAM LUPSA, DOINA TATAR

Abstract. The problem of feature selection in Word Sense Discrimination (a subtask of Word Sense Disambiguation) is crucial for the accuracy of results. The paper proposes as a new feature the length of words [1]. Some combination between this feature and other features usually used are studied and presented.

1. Introduction

The task of Word Sense Discrimination is to divide the occurrences of a word into a number of classes. It differs from the more general Word Sense Disambiguation (WSD) [20, 4] in the fact that we need only to determine which occurrences have the same meanings and not what the meaning actually is. The result is that a reference to an external knowledge source for sense definition is not required for the task of Word Sense Discrimination. Since WSD is a necessary step in a large range of applications, for many problems in information access it is sufficient to solve the discriminating problem only.

In this paper we present some remarks about feature selection in context-group discrimination method. The method was first introduced by Shutze ([15]) and consist in an unsupervised grouping of a set of contextually similar occurrences of an ambiguous word into a same cluster. The approach of this problem is based on the strong contextual hypothesis of Miller and Charles ([6]) which states that "two words are semantically related to the extent that their contextual representations are similar".

In [9] the authors systematically compare unsupervised word sense discrimination using different features of representing contexts and different clustering methods. In this paper we present some experiments made with SenseCluster using a corpus in Romanian language.

The hypothesis we introduce in this paper is that longer words carry more semantic significance on their own.

Received by the editors: November 1, 2005.

2000 *Mathematics Subject Classification.* 68T50, 68Q32.

1998 *CR Categories and Descriptors.* I.2.7 [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCE – *Natural Language Processing* .

The paper is structured as follows. Section 1 and 2 contain an introduction in Word Sense Discrimination problem and a presentation of SenseClusters tool. Section 3 presents the problem of feature selection in learning algorithms as well as a new introduced feature characterisation. Section 4 contains the experiment and evaluations of the results. In Section 5 some conclusions and future directions are presented.

2. SenseClusters

SenseClusters is a freely available word sense discrimination system ([17]) developed at the University of Minnesota, Duluth. It provides support for [9]: feature selection from an input corpus selected by user, for several different context representation methods, for various clustering algorithms and for evaluation of the discovered clusters. SenseClusters creates clusters made up of some contexts (instances) in which a given target ambiguous word occurs. A context is a group of 2 or 3 sentences, one of which contains the target word. Processing starts by selecting a corpus (in format of Senseval contests) and then selecting of features. SenseClusters supports the use of most frequent words (unigrams), the most frequent groups of two words with or without words intervening between them (bigrams) and co-occurence features (bigrams that include the target ambiguous word). The method used by the system is to represent each context as a vector. This vector could be binary, if it shows that a feature occurs or not in the context, or the frequency vector, if it shows how often the feature occurs in the context. This association of features with contexts is called “first order context vector”, as different of “second order context vectors”, introduced in [15]. There, the context vector is the average of the first order vectors associated with the words that occur in the context.

SenseClusters interface provides support for a number of clustering techniques provided by CLUTO, a Clustering Toolkit ([12]). It also offers the options for a number of similarity measures as simple matching, the cosine, the Jaccard and the Dice measures.

SenseClusters produces clusters of contexts where each cluster refers to a particular sense. The evaluation of these clusters is made with the help of an existing external knowledge of correct senses (the gold standard senses). The system produces a confusion matrix which shows the distribution of correct senses in each of the discovered clusters. The problem of assigning the maximally accurate discrimination becomes one of re-ordering the columns of the confusion matrix to maximize the diagonal sum. This method corresponds to several known methods in operation research.

3. Feature selection in learning algorithms

Notational conventions for WSD used in the following are as in [4], [20]:

- w– the ambiguous word (*target word*);
- $v_1, \cdots, v_J$— words used as contextual features for disambiguation of w.

Regarding $v_1, \cdots, v_J$, there are many possibilities. In [5] the author enumerated some sets of good indicators of word senses which could be selected as features:

- 0-param features, which can be used or not, without any parameter to set (as example the part of speech (POS) of a surrounding word). In addition in [5] are mentioned 0-param features as: verb before, verb after, noun before, noun after, named entity before, named entity after, preposition before, preposition after, pronoun before, pronoun after;
- 1-param features, which have one variable parameter that can be set to a specific value; (for example the length of a window of surrounding words or the position of a collocated word with the ambiguous word w);
- 2-param features, which have two parameters associated. As an example consider "a number" of words (the first parameter) which occur at least "a number" of times (the second parameter). As a second example consider "a number" of bigrams (first parameter) occurring at least "a number" of times (the second parameter).

A system used at contest Senseval 2 during the *English all words* task and *English lexical sample task*, based on these features selection, was ranked as the best performing one in the ranking made before the deadline.

In [7] the features are considered in the following three categories:

- morphological features (number for nouns, tense for verbs);
- POS features of two words immediately preceding and following the ambiguous word;
- collocation features which indicate if a particular word occurs in a window with the ambiguous word.

3.1. **New Word Feature Characterization.** In this paper we propose to take into consideration also the length of the words and we examinate this on the Romanian language word sense discrimination case.

Our ideea is based on two facts:

(1) The first is that longer words carry more semantic information of their own. For example, most prepositions and conjunctions are shorter words in a given language.
(2) The other is that longer words are less predilect to accumulate new meanings .

In Romanian, a special case of written word polysemy exists in the case of different words with different pronunciation and the same spelling. In what follows we will refer to it as *f-homonimy.* Because f-homonims are different words with the same written form, in what follows we need to make the distinction between the word and the vector of letters that constitutes the written form of the word.

Whenever we need to make this distinction, we will use the term “word form” to refere the letters of the written form of the word.

In order to test our intuition (item (2) in the list above), for each possible word length we compared the number of word forms in Romanian language with the number of word forms which have f-homonims. In this paper, we study the case of f-homonimy instead of polisemy for Romanian language, because a free version of DEX ([2]) is available and because distinct words with the same form (f-homonims) are easy to identify because they have distinct entries in DEX.

By using the Romanian DEX ([2]) we have extracted all distinct word forms that are entries in DEX and also all word forms that have more than one entry in DEX. Their absolute frequencies for each possible word length is depicted on the same graphic (figure 1).

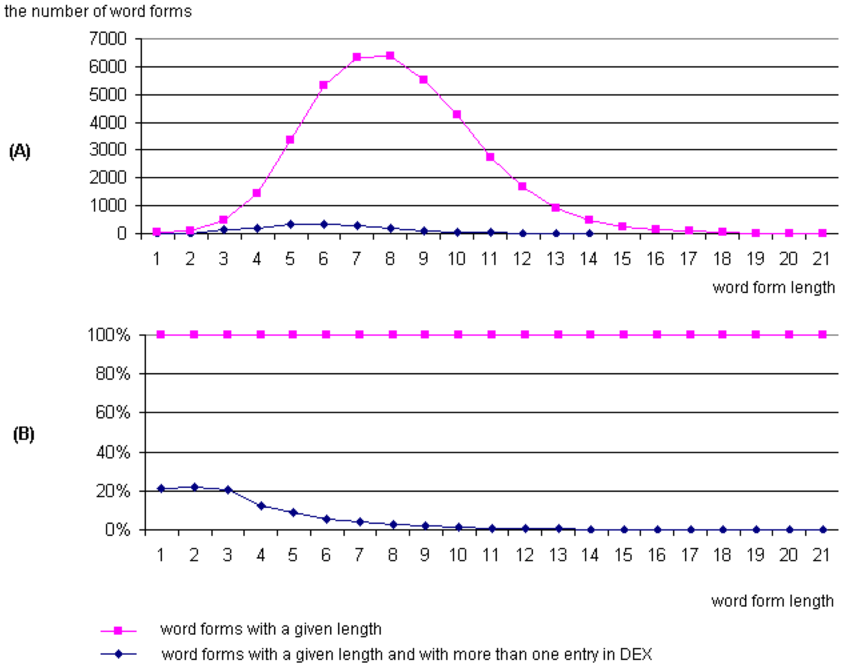


FIGURE 1. The number of F-Homonymic word lenghts compared with lengths of all words for Romanian language

Sub-Figure (1.A) presents the histogram of the absolute word form frequency for a given length and absolute frequency of f-homonimic word forms. Sub-Figure (1.B) represents the ratio between frequency of f-homonimic word forms and frequency of all word forms for each possible word length. The representation from

(1.B) indicates more visibly that the ratio between f-homonimic word forms frequency and word forms frequency is decreasing while word length is increasing. This is the remark (related to 2) we intend to take advantage of when we introduce the feature selection parametrized by the length of the word feature.

4. The Experiment

4.1. **How we evaluate our hypothesis.** The hypothesis we introduce in this paper is that longer words carry more semantic significance on their own and are less polysemantic. A consequence of this is that selecting longer words as attributes of a context should characterize better the context.

One way of doing word sense discrimination is to cluster the contexts of ambiguous word. A cluster of contexts corresponds to contexts of the same meaning of the given word. Choosing better attributes for the contexts should bring better results for the word sense discrimination process.

We evaluate the importance and the influence on clustering process by selecting different word lengths as context attributes and we use that for clustering contexts of some polisemous Romanian words. For clustering contexts of word occurences, we use the SenseClusters program. A presentation of the application was made by its authors in [3], [13], [12]. The use of clustering similar contexts in word sense discrimination and the influence of different parameters is studied in [9], [11],[10].

We want to represent the meaning of a context as an average meaning of the words that appear in the context. Following this idea, we choose to use the agglomerative clustering method with average link criteria function. There are argues in literature [11], [10] that the average link criteria function fares well.

We use the unigram and bigram type of feature. There is not a consentaneous opinion about which of them is better. The work presented in [8] emphasizes the importance of bigram in word sense dezambiguation, while in [11] co-occurences and unigrams achieved the overall best results. For our data, unigrams performed better than bigrams.

From SenseClusters point of view, bigrams features are pairs of words that occur in a given order within some distances from each other. We choose a window size of three, meaning that there could be at most one intervening word between the first and the second word that make a bigram.

Unigrams are single words that occur in the same context as the target word and they are made up of all the words found in context.

The new parameter we introduce is the length of feature words. We used as word length parameter the values 2, 3, ..., 10. In general, the length parameter indicates that the length of the word feature is greater than the parameter value. For unigrams, it means that selected feature words are longer than the indicated value. In the case of bigram attributes, this parameter refers to the two feature words which are selected according to the bigram model. In this case we enforce that both feature words to be longer than the length parameter.

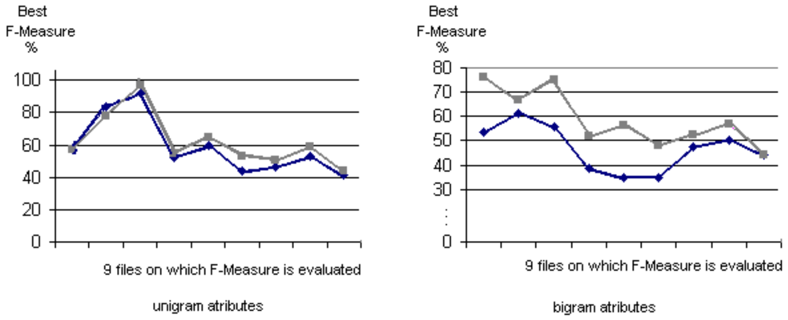


Figure 2. Best values of FMeasure (in percent) with and without word length filter (on the gray line - with word length filter; on the black line - without word length filter)

For each dataset and the two type of attribute we have worked with, we compared the best result obtained with using the word length parameter and without selecting attributes based on their word length. The results are presented in table 4.1. The dark color represents the best results obtained without any restriction of word length and the light color represents the results obtained by using the word length restriction.

That type of general results obtained by the application with and without using word length parameter justifies the fact the word length parameter is worthy to be studied. In the next section we are going to present in detail the results for all the combinations of parameters we have studied.

4.2. **Data.** We use as test data a Romanian corpus from SenseEval [1] that is not annotated with POS information. It contains contexts for 39 words, among them there are 25 nouns. There are about 1 million words and 7674 contexts. The file contains 248 number of senses to be disambiguated/discriminated.

We choose the words *actiune* (*action*), *eruptie* (*eruption*), *problema* (*problem*). For each of them we have selected three set of contexts, as follows:

- one is formed by all contexts of the word and its senses in the original selected SenseEval file;
- the other two sets of contexts are built by dividing the corpus into two parts, with almost the same size.

The characteristics of the chosen contexts are presented in table 1.

4.3. **Evaluation Method.** In this study we use all the 9 datasets presented in subsection 4.2 for each word length parameter value in $\{2, 3, \ldots, 10\}$. We evaluated each result by using F-measure values computed by SenseClusters.

[1] We used the file RomanianLS.unlabeled

word	number of		
	senses	contexts	words in contexts
actiune	8	299	about 50000
actiune	7	138	
actiune	8	161	
eruptie	2	54	about 8500
eruptie	2	18	
eruptie	2	36	
problema	6	288	about 45000
problema	5	142	
problema	5	146	

Table 1. Characteristics of the sets of word contexts

For overall evaluation we used two methods. One is by computing the average of the F-measure values of all datasets for each word length parameter (figures 3(a) and 4(a)). The other is based on ranking the F-measure values and we present it in what follows.

Borrowing ideas from the notion of Pareto dominance ([18], [19]) we define the dominance number (definition 2) and use it for a second type of evaluation of the importance of word length parameter.

Definition 1 (Better solution)**.** *Let $\mathcal{S}$ be solution space, $x, y \in \mathcal{S}$ and $eval : \mathcal{S} \to \Re$ a solution evaluation function. We define:*

$$better\_solution(x; y) = \begin{cases} 1 & if\ eval(x) \geq eval(y) \\ 0 & if\ eval(x) \leq eval(y) \end{cases}$$

In other words, the definition (1) says that $better\_solution(x; y) = 1$ iff x is a better solution than y; otherwise $better\_solution(x; y) = 0$.

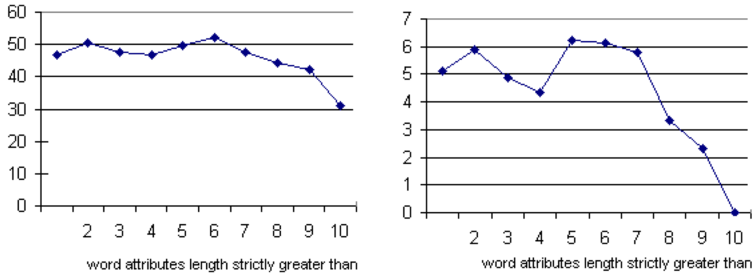
Definition 2 (Dominance number)**.** *Let $\mathcal{S}$ be solution space , $y, x_1, x_2, \ldots x_n \in \mathcal{S}$ and $eval : \mathcal{S} \to \Re$ a solution evaluation function. The dominance number of y over $x_1, x_2, \ldots x_n$ is:*

$$dominance\_number(y; x_1, x_2, \ldots x_n) = \sum_{i=1}^{n} better\_solution(y, x_i)$$

Let us consider a data set and the F-measure value for each parameter value. We computed the dominace number (definition 2) for each parameter value. The overall evaluation of each parameter value is made by averaging the dominace number for all data sets considered.

4.4. **Results.** We do that independently for bigram and unigram feature and for each word length parameter. The results for bigram feature is presented in figure 3 and the results for unigram feature are presented in figure 4. On x axis, graphic representations contain the results for each value used as length parameter. The y axis correspond to the average of the evaluation measure.

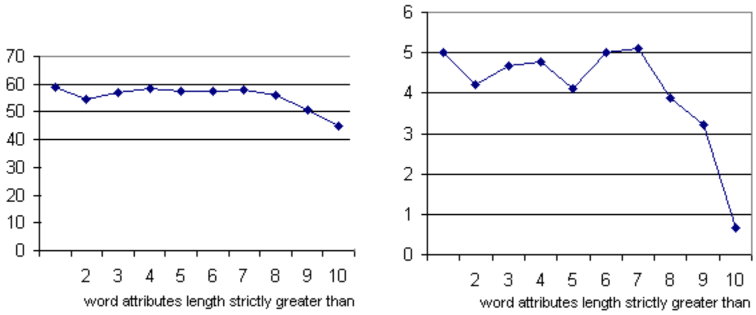
The use of the average of dominance number values and the results are given in figures 3(b) and 4(b).



(a) average of FMeasures

(b) average of dominace numbers corresponding to FMeasures

FIGURE 3. Evaluation for bigram word features



(a) average of FMeasures

(b) average of dominace numbers corresponding to FMeasures

FIGURE 4. Evaluation for unigram word features

It can easily be observed that the best results for bigram type of features are achieved for length parameter with value six if we are giving credit to the evaluation that use average of F-measure values (figure 3(a)). If we are using the evaluation technique based on dominance number (figure 3) we should say that we get better results if we select the words longer than five characters. As the two different evaluation techniques get slightly different results, we could only claim that best results are obtained for length parameter with value five or six.

When using unigrams, first of the two evaluation methods indicates as best results those obtained without using the length parameter (or choosing the value 0 for this parameter), closely followed by results obtained for length parameter with value four (figure 4(a)). The second evaluation indicates selection of word feature by the length greater than seven as getting best results (figure 4(b)).

5. Conclusions

The experiments confirm the intuition that the results are better when using longer words as features. The length parameter value for which the results are better, cover the set $\{4, 5, 6, 7\}$. The non-achievement of this study is that we couldn't indicate a unique best value for the word length parameter.

Some future studies about other features which could improve word sense discrimination results and the possibilities to integrate these conclusions in existing WSD methods are in this moment in our attention.

References

[1] D. Avram: *Extragerea informatiilor de tip semantic din texte folosind clasificare nesupervizata*. PhD. Thesis, Babes-Bolyai University Cluj-Napoca (in Romanian)(in preparation)

[2] *dexonline - Romanian **EX**planatory **D**ictionary* , 2004.
http://www.dexonline.ro/

[3] A. Kulkarni, T. Pedersen: *SenseClusters: Unsupervised Clustering and Labeling of Similar Contexts*. Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 105–108, 2005
http://www.aclweb.org/anthology/P/P05/P05-3027

[4] C. Manning, H. Schutze: *Foundation of statistical natural language processing*. MIT, 1999

[5] R. Mihalcea: *Word Sense Disambiguation with pattern learning and automatic feature selection*. Natural Language Engineering, 1:1–15, 2002

[6] G.Miller, W.G. Charles: *Contextual corelates of semantic similarity*. Language and Cognitive Processes, 6:1–28, 1991

[7] T. Pedersen, R. Bruce, J. Wiebe: *Sequential Model Selection for Word Sense Disambiguation*. Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97), 1997
http://www.d.umn.edu/~tpederse/Pubs/anlp97.ps

[8] T. Pedersen: *A decision tree of bigrams is an accurate predictor of word sense*. In Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01), 2001
http://www.d.umn.edu/~tpederse/Pubs/naacl01.pdf

[9] A. Purandare, T. Pedersen: *SenseClusters - Finding Clusters that Represent Word Senses*. In Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI), 2004
http://www.d.umn.edu/~tpederse/Pubs/AAAI04PurandareA.pdf

[10] A. Purandare, T. Pedersen: *Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces*. Proceedings of the Conference on Computational Natural Language

Learning (CoNLL) 2004
http://www.d.umn.edu/∼tpederse/Pubs/conll04-purandarep.pdf

[11] A. Purandare: *Unsupervised Word Sense Discrimination by Clustering Similar Context.* PhD. Thesis, University of Minnesota, 2004

[12] T. Pedersen and A. Kulkarni: *Identifying Similar Words and Contexts in Natural Language with SenseClusters.* Proceedings of the Twentieth National Conference on Artificial Intelligence, 2005
http://www.d.umn.edu/∼tpederse/Pubs/aaai2005-demo-sc.pdf

[13] T. Pedersen: *Language Independent Methods of Clustering Similar Contexts.* Eurolan 2005 Summer School. Tutorials, 2005

[14] H. Schutze: *Automatic Word Sense Discrimination.* Computational linguistics, 24(1):97–123, 1998

[15] H. Schutze: *Dimensions of meaning.* Proceedings of Supercomputing '92, pp.787–796, 1992

[16]

[17] T. Pedersen, A. Purandare, A. Kulkarni: *Senseclusters home page.* 2005.
http://senseclusters.sourceforge.net/

[18] W. Stadler: *A Survey of Multicriteria Optimization, or the Vector Maximum Problem.* Journal of Optimization Theory and Applications 29:1–52, 1979

[19] R.E. Steuer: *Multiple Criteria Optimization: Theory, Computation and Application.* New York: Wiley, 1986

[20] D. Tatar: *Word sense disambiguation; a short survey.* Studia Univ. Babes-Bolyai, XLIX(2):17–27, 2004

Babes-Bolyai University, Faculty of Mathematics and Computer Science, Department of Computer Science, Cluj-Napoca, Romania

E-mail address: davram@cs.ubbcluj.ro, dtatar@cs.ubbcluj.ro