Improving Name Discrimination: A Language Salad Approach

Ted Pedersen and **Anagha Kulkarni**
Department of Computer Science
University of Minnesota, Duluth
Duluth, MN 55812 USA
{tpederse,kulka020}@d.umn.edu

Roxana Angheluta
Attentio SA
B-1030 Brussels, Belgium
roxana@attentio.com

Zornitsa Kozareva
Dept. de Lenguajes y Sistemas Informáticos
University of Alicante
03080 Alicante, Spain
zkozareva@dlsi.ua.es

Thamar Solorio
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79902 USA
tsolorio@utep.edu

Abstract

This paper describes a method of discriminating ambiguous names that relies upon features found in corpora of a more abundant language. In particular, we discriminate ambiguous names in Bulgarian, Romanian, and Spanish corpora using information derived from much larger quantities of English data. We also mix together occurrences of the ambiguous name found in English with the occurrences of the name in the language in which we are trying to discriminate. We refer to this as a language salad, and find that it often results in even better performance than when only using English or the language itself as the source of information for discrimination.

1 Introduction

Name ambiguity is a problem that is increasing in complexity and scope as online information sources grow and expand their coverage. Like words, names are often ambiguous and can refer to multiple underlying entities or concepts. Web searches for names can often return results associated with multiple people or organizations in a disorganized and unclear fashion. For example, the top 10 results of a Google search for *George Miller* includes a mixture of entries for two different entities, one a psychology professor from Princeton University and the other the director of the film *Mad Max*.[1]

Name discrimination takes some number of contexts that include an ambiguous name, and divides them into groups or clusters, where the contexts in each cluster should ideally refer to the same underlying entity (and each cluster should refer to a different entity). Thus, if we are given 10,000 contexts that include the name *John Smith*, we would want to divide those contexts into clusters corresponding to each of the different underlying entities that share that name.

We have developed an unsupervised method of name discrimination (Pedersen et al., 2005). We have shown the method to be language independent (Pedersen et al., 2006), which is to say we can apply it to English contexts as easily as we can apply it to Romanian or French. However, we have observed that there are situations where the number of contexts in which an ambiguous name occurs is relatively small, perhaps because the name itself is unusual, or because the quantity of data available for language is limited in general. These problems of scarcity can make it difficult to apply these methods and discriminate ambiguous names, especially in languages with fewer online resources.

This paper presents a method of name discrimination is based on using a larger number of contexts in English that include an ambiguous name, and applying information derived from these contexts to the discrimination of that name in another language, where there are many fewer contexts. We also show that mixing English contexts with the contexts to be discriminated can result in a performance improvement over only using the English or the original contexts alone.

2 Discrimination by Clustering Contexts

Our method of name discrimination is described in more detail in (Pedersen et al., 2005), but in general is based on an unsupervised approach to word sense discrimination introduced by (Purandare and

[1] Search conducted January 4, 2006.

Pedersen, 2004), which builds upon earlier work in word sense discrimination, including (Schütze, 1998) and (Pedersen and Bruce, 1997).

Our method treats each occurrence of an ambiguous name as a context that is to be clustered with other contexts that also include the same name. In this paper, each context consists of about 50 words, where the ambiguous name is generally in the middle of the context. The goal is to cluster similar contexts together, based on the presumption that the occurrences of a name that appear in similar contexts will refer to the same underlying entity. This approach is motivated by both the *distributional hypothesis* (Harris, 1968) and the *strong contextual hypothesis* (Miller and Charles, 1991).

2.1 Feature Selection

The contexts to be clustered are represented by lexical features which may be selected from either the contexts being clustered, or from a separate corpus. In this paper we use both approaches. We cluster the contexts based on features identified in those very same contexts, and we also cluster the contexts based on features identified in a separate set of data (in this case English). We explore the use of a mixed feature selection strategy where we identify features both from the data to be clustered and the separate corpus of English text. Thus, our *feature selection data* may come from one of three sources: the contexts to be clustered (which we will refer to as the *evaluation contexts*), English contexts which include the same name but are not to be clustered, and the combination of these two (our so-called Language Salad or Mix).

The lexical features we employ are bigrams, that is consecutive words that occur together in the corpora from which we are identifying features. In this work we identify bigram features using Pointwise Mutual Information (PMI). This is defined as the log of the ratio of the observed frequency with which the two words occur together in the feature selection data, to the expected number of times the two words would occur together in a corpus if they were independent. This expected value is estimated simply by taking the product of the number of times the two words occur individually, and dividing this by the total number of bigrams in the feature selection data. Thus, larger values of PMI indicate that the observed frequency of the bigram is greater than would be expected if the two words were independent.

In these experiments we take the top 500 ranked bigrams that occur five or more times in the feature selection data. We also exclude any bigram from consideration that is made up of one or two stop words, which are high frequency function words that have been specified in a manually created list. Note that with smaller numbers of contexts (usually 200 or fewer), we lower the frequency threshold to two or more.

In general PMI is known to have a bias towards pairs of words (bigrams) that occur a small number of times and only with each other. In this work that is a desirable quality, since that will tend to identify pairs of words that are very strongly associated with each other and also provide unique discriminating information.

2.2 Context Representation

Once the bigram features have been identified, then the contexts to be clustered are represented using *second order co-occurrences* that are derived from those bigrams. In general a second order co-occurrence is a pair of words that may not occur with each other, but that both occur frequently with a third word. For example, *garden* and *fire* may not occur together often, but both commonly occur with *hose*. Thus, *garden hose* and *fire hose* represent first order co–occurrences, and *garden* and *fire* represent a second order co–occurrence.

The process of creating the second order representation has several steps. First, the bigram features identified by PMI (the top ranked 500 bigrams that have occurred 5 or more times in the feature selection data) are used to create a word by word co–occurrence matrix. The first word in each bigram represents a row in the matrix, and the second word in each bigram represents a column. The cells in the matrix contain the PMI scores. Note that this matrix is not symmetric, and that there are many words that only occur in either a row or a column (and not both) because they tend to occur as the first or second word in a bigram. For example, *President* might tend to be a first word in a bigram (e.g., *President Clinton*, *President Putin*), whereas last names will tend to be the second word.

Once the co–occurrence matrix is created, then the contexts to be clustered can be represented. Each word in the context is checked to see if it

has a corresponding row (i.e., vector) in the co–occurrence matrix. If it does, that word is replaced in the context by the row from the matrix, so that the word in the context is now represented by the vector of words with which it occurred in the feature selection data. If a word does not have a corresponding entry in the co–occurrence matrix, then it is simply removed from the context. After all the words in the context are checked, then all of the vectors that are selected are averaged together to create a vector representation of the context.

Then these contexts are clustered into a pre–specified number of clusters using the k–means algorithm. Note that we are currently developing methods to automatically select the number of clusters in the data (e.g., (Pedersen and Kulkarni, 2006)), although we have not yet applied them to this particular work.

3 The Language Salad

In this paper, we explore the creation of a second order representation for a set of evaluation contexts using three different sets of feature selection data. The co–occurrence matrix may be derived from the evaluation contexts themselves, or from a separate set of contexts in a different language, or from the combination of these two (the Salad or Mix).

For example, suppose we have 100 Romanian evaluation contexts that include an ambiguous name, and that same name also occurs 10,000 times in an English language corpus.[2] Our goal is to cluster the 100 Romanian contexts, which contain all the information that we have about the name in Romanian. While we could derive a second order representation of the contexts, the resulting co–occurrence matrix would likely be very small and sparse, and insufficient for making good discrimination decisions. We could instead rely on first order features, that is look for frequent words or bigrams that occur in the evaluation contexts, and try and find evaluation contexts that share some of the same words or phrases, and cluster them based on this type of information. However, again, the small number of contexts available would likely result in very sparse representations for the contexts, and unreliable clustering results.

Thus, our method is to derive a co–occurrence matrix from a language for which we have many occurrences of the ambiguous name, and then use that co–occurrence matrix to represent the evaluation contexts. This relies on the fact that the evaluation contexts will contain at least a few names or words that are also used in the larger corpus (in this case English). In general, we have found that while this is not always true, it is often the case.

We have also experimented with combining the English contexts with the evaluation contexts, and building a co–occurrence matrix based on this combined or mixed collection of contexts. This is the language salad that we refer to, a mixture of contexts in two different languages that are used to derive a representation of the evaluation contexts.

4 Experimental Data

We use data in four languages in these experiments, Bulgarian, English, Romanian, and Spanish.

4.1 Raw Corpora

The Romanian data comes from the 2004 archives of the newspaper Adevarul (The Truth)[3]. This is a daily newspaper that is among the most popular in Romania. While Romanian normally has diacritical markings, this particular newspaper does not include those in their online edition, so the alphabet used was the same as English.

The Bulgarian data is from the Sega 2002 news corpus, which was originally prepared for the CLEF competition.[4] This is a corpus of news articles from the Newspaper Sega[5], which is based in Sofia, Bulgaria. The Bulgarian text was transliterated (phonetically) from Cyrillic to the Roman alphabet. Thus, the alphabet used was the same as English, although the phonetic transliteration leads to fewer cognates and borrowed English words that are spelled exactly the same as in English text.

The Spanish corpora comes from the Spanish news agency EFE from the year 1994 and 1995. This collection was used in the Question Answering Track at CLEF-2003, and also for CLEF-2005. This text is represented in Latin-1, and includes the usual accents that appear in Spanish.

The English data comes from the GigaWord corpus (2nd edition) that is distributed by the Linguistic Data Consortium. This consists of more

[2] We assume that the names either have the same spelling in both languages, or that translations are readily available.

[3] http://www.adevarulonline.ro/arhiva

[4] http://www.clef-campaign.org

[5] http://www.segabg.com

than 2 billion words of newspaper text that comes from five different news sources between the years 1994 and 2004. In fact, we subdivide the English data into three different corpora, where one is from 2004, another from 2002, and the third from 1994-95, so that for each of the evaluation languages (Bulgarian, Spanish, and Romanian) we have an English corpus from the same time period.

4.2 Evaluation Contexts

Our experimental data consists of evaluation contexts derived from the Bulgarian, Romanian, and Spanish corpora mentioned above. We also have English corpora that includes the same ambiguous names as found in the evaluation contexts.

In order to quickly generate a large volume of experimental data, we created evaluation contexts from the corpora for each of our four languages by conflating together pairs of well known names or places, and that are generally not highly ambiguous (although some might be rather general). For example, one of the pairs of names we conflate is *George Bush* and *Tony Blair*. To do that, every occurrence of both of these names is converted to an ambiguous form (GB_TB, for example), and the discrimination task is to cluster these contexts such that their original and correct name is re–discovered. We retain a record of the original name for each occurrence, so as to evaluate the results of our method. Of course we do not use this information anywhere in the process outside of evaluation.

The following pairs of names were conflated in all four of the languages: George Bush-Tony Blair, Mexico-India, USA-Paris, Ronaldo-David Beckham (2002 and 2004), Diego Maradona-Roberto Baggio (1994-95 only), and NATO-USA. Note that some of these names have different spellings in some of our languages, so we look for and conflate the native spelling of the names in the different language corpora. These pairs were selected because they occur in all four of our languages, and they represent name distinctions that are commonly of interest, that is they represent ambiguity in names of people and places. With these pairs we are also following (Nakov and Hearst, 2003) who suggest that if one is introducing ambiguity by creating *pseudo—words* or conflating names, then these words should be related in some way (in order to avoid the creation of very sharp or obvious sense distinctions).

4.3 Discussion

For each of the three evaluation languages (Bulgarian, Romanian, and Spanish) we have contexts for five different name conflate pairs that we wish to discriminate. We have corresponding English contexts for each evaluation language, where the dates of both are approximately the same. This temporal consistency between the evaluation language and English is important because the contexts in which a name is used may change over time. In 1994, for example, Tony Blair was not yet Prime Minister of England (he became PM in 1997), and references to George Bush most likely refer to the US President who served from 1988 until 1992, rather than the current US President (who began his term in office in 2001). In 1994 the current (as of 2006) US President had just been elected governor of Texas, and was not yet a national figure. This points out that George Bush is an example of an ambiguous name, but our observation has been that in the 2002 and 2004 data (Romanian and Bulgarian) nearly all occurrences are associated with the current president, and that most of the occurrences in 1994-95 (Spanish) refer to the former US President. This illustrates an important point: it is necessary to consider the perspective represented by the different corpora. There is little reason to expect that news articles from Spain in 1994 and 1995 would focus much attention on the newly elected governor of Texas in the United States.

Tables 1, 2, and 3 show the number of contexts that have been collected for each name conflate pair. For example, in Table 1 we see that there are 746 Bulgarian contexts that refer to either Mexico or India, and that of these 51.47% truly refer to Mexico, and 48.53% to India. There are 149,432 English contexts that mention Mexico or India, and the Mix value shown is simply the sum of the number of Bulgarian and English contexts.

In general these tables show that the English contexts are much larger in number, however, there are a few exceptions with the Spanish data. This is because the EFE corpus is relatively large as compared to the Bulgarian and Romanian corpora, and provides frequency counts that are in some cases comparable to those in the English corpus.

5 Experimental Methodology

For each of the three evaluation languages (Bulgarian, Romanian, Spanish) there are five name conflate pairs. The same name conflate pairs are used for all three languages, except for Diego Maradona-Roberto Baggio which is only used with Spanish, and Ronaldo-David Beckham, which is only used with Bulgarian and Romanian. This is due to the fact that in 1994-95 (the era of the Spanish data) neither Ronaldo nor David Beckham were as famous as they later became, so they were mentioned somewhat less often than in the 2002 and 2004 corpora. The other four name conflate pairs are used in all of the languages.

For each name conflate pair we create a second order representation using three different sources of features selection data: the evaluation contexts themselves, the corresponding English contexts, and then the mix of the evaluation contexts and the English contexts (the Mix). The objective of these experiments is to determine which of these sources of feature selection data results in the highest F-Measure, which is the harmonic mean of the precision and recall of an experiment.

The precision of each experiment is the number of evaluation contexts clustered correctly, divided by the number of contexts that are clustered. The clustering algorithm may choose not to assign every context to a cluster, which is why that denominator may not be the same as the number of evaluation contexts. The recall of each experiment is the the number of correctly clustered evaluation contexts divided by the total number of evaluation contexts. Note that for each of the three variations for each name conflate pair experiment exactly the same evaluation language contexts are being discriminated, all that is changing in each experiment is the source of the feature selection data. Thus the F-measures for a name conflate pair in a particular language can be compared directly. Note however that the F-measures across languages are harder to compare directly, since different evaluation contexts are used, and different English contexts are used as well.

There is a simple baseline that can be used as a point of comparison, and that is to place all of the contexts for each name conflate pair into one cluster, and say that there is no ambiguity. If that is done, then the resulting F-Measure will be equal to the majority percentage of the true underlying entity as shown in Tables 1, 2, and 3. For example, for Bulgarian, if the 746 Bulgarian contexts for Mexico and India are all put into the same cluster, the resulting F-Measure would be 51.47%, because we would simply assign all the contexts in the cluster to the more common of the two entities, which is Mexico in this case.

6 Experimental Results

Tables 1, 2, and 3 show the results for our experiments, language by language. Each table shows the results for the 15 experiments done for each language: five name conflate pairs, each with three different sources of feature selection data. The row labeled with the name of the evaluation language reports the F-Measure for the evaluation contexts (whose number of occurrences is shown in the far right column) when the feature selection data is the evaluation contexts themselves. The rows labeled English and Mix report the F-Measures obtained for the evaluation contexts when the feature selection data is the English contexts, or the Mix of the English and evaluation contexts.

6.1 Bulgarian Results

The Bulgarian results are shown in Table 1. Note that the number of contexts for English is considerably larger than for Bulgarian for all five name conflate pairs. The Bulgarian and English data came from 2002 news reports.

The Mix of feature selection data results in the best performance for three of the five name conflate pairs: George Bush - Tony Blair, Ronaldo - David Beckham, and NATO - USA. For remaining two name conflate pairs, just using the Bulgarian evaluation contexts results in the highest F-Measure (Mexico-India, USA-Paris).

We believe that this may be partially due to the fact that the two cases where Bulgarian leads to the best results are for very general or generic underlying entities: Mexico and India, and then the USA and Paris. In both cases, contexts that mention these entities could be discussing a wide range of topics, and the larger volumes of English data may simply overwhelm the process with a huge number of second order features. In addition, it may be that the English and Bulgarian corpora contain different content that reflects the different interests of the original readership of this text. For example, news that is reported about India might be rather different in the United States (the source of most

Table 1: Bulgarian Results (2002): Feature Selection Data, F-Measure, and Number of Contexts

George Bush (73.43) - Tony Blair (26.57)		
Mix	68.37	11,570
Bulgarian	55.78	651
English	36.15	10,919
Mexico (51.47) - India (48.53)		
Bulgarian	70.97	746
Mix	55.01	150,178
English	48.15	149,432
USA (79.53) - Paris (20.47)		
Bulgarian	58.67	3,283
Mix	51.68	56,044
English	49.66	52,761
Ronaldo (61.25) - David Beckham (38.75)		
Mix	64.88	8,649
Bulgarian	52.75	320
English	48.11	8,329
NATO (87.37) - USA (12.63)		
Mix	75.44	54,193
Bulgarian	65.92	3,770
English	60.44	50,423

Table 2: Romanian Results (2004): Feature Selection Data, F-Measure, and Number of Contexts

Tony Blair (72.00) - George Bush (28.00)		
English	64.23	11,616
Mix	54.31	11,816
Romanian	50.75	200
India (53.66) - Mexico (46.34)		
Romanian	50.93	82
English	47.30	88,247
Mix	42.55	88,329
USA (60.29) - Paris (39.71)		
English	59.05	45,346
Romanian	58.76	700
Mix	57.91	46,046
David Beckham (55.56) - Ronaldo (44.44)		
Mix	81.00	4,365
English	70.85	4,203
Romanian	52.47	162
NATO (58.05) - USA (41.95)		
Mix	60.48	43,508
Romanian	51.20	1,168
English	38.91	42,340

of the English data) than in Bulgaria. Thus, the use of the English corpora might not have been as helpful in those cases where the names to be discriminated are more global figures. For example, Tony Blair and George Bush are probably in the news in the USA and Bulgaria for many of the same reasons, thus the underlying content is more comparable than that of the more general entities (like Mexico and India) that might have much different content associated with them.

We observed that Bulgarian tends to have fewer cognates or shared names with English than do Romanian and English. This is due to the fact that the Bulgarian text is transliterated. This may account for the fact that the English-only results for Bulgarian are very poor, and it is only in combination with the Bulgarian contexts that the English contexts show any positive effect. This suggests that there are only a few words in the Bulgarian contexts that also occur in English, but those that do have a positive impact on clustering performance.

6.2 Romanian Results

The Romanian results are shown in Table 2. The Romanian and English contexts come from 2004.

The Mix of Romanian and English contexts for feature selection results in improvements for two of the five pairs (David Beckham - Ronaldo, and NATO - USA). The use of English contexts only provides the best results for two other pairs (Tony Blair - George Bush, and USA - Paris, although in the latter case the difference in the F-Measures that result from the three sources of data is minimal). There is one case (Mexico-India) where using the Romanian contexts as feature selection data results in a slightly better F-measure than when using English contexts.

The improvement that the Mix shows for David Beckham-Ronaldo is significant, and is perhaps due to fact that in both English and Romanian text, the content about Beckham and Ronaldo is similar, making it more likely that the mix of English and Romanian contexts will be helpful. However, it is also true that the Mix results in a significant improvement for NATO-USA, and it seems likely that the local perspective in Romania and the USA would be somewhat different on these two entities. However, NATO-USA has a relatively large number of contexts in Romanian as well as English, so perhaps the difference in perspective had less of an impact in those cases where the number of Ro-

Table 3: Spanish Results (1994-95): Feature Selection Data, F-Measure, and Number of Contexts

George Bush (75.58) - Tony Blair (24.42)		
Mix	78.59	2,353
Spanish	64.45	1,163
English	54.29	1,190
D. Maradona (51.55) - R. Baggio (48.45)		
English	67.65	1,588
Mix	61.35	3,594
Spanish	60.70	2,006
India (92.34) - Mexico (7.66)		
English	72.76	19,540
Spanish	66.57	2,377
Mix	61.54	21,917
USA (62.30) - Paris (37.70)		
Spanish	69.31	1,000
English	64.30	17,344
Mix	59.40	18,344
NATO (63.86) - USA (36.14)		
Spanish	62.04	2,172
Mix	58.47	27,426
English	56.00	25,254

manian contexts is much smaller (as is the case for Beckham and Ronaldo).

6.3 Spanish Results

The Spanish results are shown in Table 3. The Spanish and English contexts come from 1994-1995, which puts them in a slightly different historical era than the Bulgarian and Romanian corpora.

Due to this temporal difference, we used Diego Maradona and Roberto Baggio as a conflated pair, rather than David Beckham and Ronaldo, who were much younger and somewhat less famous at that time. Also, Ronaldo is a highly ambiguous name in Spanish, as it is a very common first name. This is true in English text as well, although casual inspection of the English text from 2002 and 2004 (where the Ronaldo-Beckham pair was included experimentally) reveals that Ronaldo the soccer player tends to occur more so than any other single entity named Ronaldo, so while there is a bit more noise for Ronaldo, there is not really a significant ambiguity.

For the Spanish results we only note one pair (George Bush - Tony Blair) where the Mix of English and Spanish results in the best performance. This again suggests that the perspective of the Spanish and English corpora were similar with respect to these entities, and their combination was helpful. In two other cases (Maradona-Baggio, India-Mexico) English only contexts achieve the highest F-Measure, and then in the two remaining cases (USA-Paris, NATO-USA) the Spanish contexts are the best source of features.

Note that for Spanish we have reasonably large numbers of contexts (as compared to Bulgarian and Romanian). Given that, it is especially interesting that English-only contexts are the most effective in two of five cases. This suggests that this approach may have merit even when the evaluation language does not suffer from problems of extreme scarcity. It may simply be that the information in the English corpora provides more discriminating information than does the Spanish, and that it is somewhat different in content than the Spanish, otherwise we would expect the Mix of English and Spanish contexts to do better than being most accurate for just one of five cases.

7 Discussion

Of the 15 name conflate experiments (five pairs, three languages), in only five cases did the use of the evaluation contexts as a source of feature selection data result in better F-Measure scores than did either using the English contexts alone or as a Mix with the evaluation language contexts. Thus, we conclude that there is a clear benefit to using feature selection data that comes from a different language than the one for which discrimination is being performed.

We believe that this is due to the volume of the English data, as well as to the nature of the name discrimination task. For example, a person is often best described or identified by observing the people he or she tends to associate with, or the places he or she visits, or the companies with which he or she does business. If we observe that *George Miller* and *Mel Gibson* occur together, then it seems we can safely infer that *George Miller* the movie director is being referred to, rather than *George Miller* the psychologist and father of WordNet.

This argument might suggest that first order co–occurrences would be sufficient to discriminate among the names. That is, simply group the evaluation contexts based on the features that occur within them, and essentially cluster evaluation

contexts based on the number of features they have in common with other evaluation contexts. In fact, results on word sense discrimination (Purandare and Pedersen, 2004) suggest that first order representations are more effective with larger number of context than second order methods. However, we see examples in these results that suggests this may not always be the case. In the Bulgarian results, the largest number of Bulgarian contexts are for NATO-USA, but the Mix performs quite a bit better than Bulgarian only. In the case of Romanian, again NATO-USA has the largest number of contexts, but the Mix still does better than Romanian only. And in Spanish, Mexico-India has the largest number of contexts and English-only does better. Thus, even in cases where we have an abundant number of evaluation contexts, the indirect nature of the second order representation provides some added benefit.

We believe that the perspective of the news organizations providing the corpora certainly has an impact on the results. For example, in Romanian, the news about David Beckham and Ronaldo is probably much the same as in the United States. These are international figures that are both external to countries where the news originates, and there is no reason to suppose there would be a unique local perspective represented by any of the news sources. The only difference among them might be in the number of contexts available. In this situation, the addition of the English contexts may provide enough additional information to improve discrimination performance in another language.

For example, in the 162 Romanian contexts for Ronaldo-Beckham, there is one occurrence of *Posh*, which was the stage name of Beckham's wife Victoria. This is below our frequency cutoff threshold for feature selection, so it would be discarded when using Romanian–only contexts. However, in the English contexts Posh is mentioned 6 times, and is included as a feature. Thus, the one occurrence of *Posh* in the Romanian corpus can be well represented by information found in the English contexts, thus allowing that Romanian context to be correctly discriminated.

8 Conclusions

This paper shows that a method of name discrimination based on second order context representations can take advantage of English contexts, and the mix of English and evaluation contexts, in order to perform more accurate name discrimination.

9 Acknowledgments

This research is supported by a National Science Foundation Faculty Early CAREER Development Award (#0092784). All of the experiments in this paper were carried out with version 0.71 SenseClusters package, which is freely available from http://senseclusters.sourceforge.net.

References

Z. Harris. 1968. *Mathematical Structures of Language*. Wiley, New York.

G.A. Miller and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

P. Nakov and M. Hearst. 2003. Category-based pseudowords. In *Companion Volume to the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 67–69, Edmonton, Alberta, Canada, May 27 - June 1.

T. Pedersen and R. Bruce. 1997. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 197–207, Providence, RI, August.

T. Pedersen and A. Kulkarni. 2006. Selecting the ïightnïumber of senses based on clustering criterion functions. In *Proceedings of the Posters and Demo Program of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April.

T. Pedersen, A. Purandare, and A. Kulkarni. 2005. Name discrimination by clustering similar contexts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 220–231, Mexico City, February.

T. Pedersen, A. Kulkarni, R. Angheluta, Z. Kozareva, and T. Solorio. 2006. An unsupervised language independent method of name discrimination using second order co-occurrence features. In *Proceedings of the Seventh International Conference on Intelligent Text Processing and Computational Linguistics*, pages 208–222, Mexico City, February.

A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 41–48, Boston, MA.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.