An Unsupervised Language Independent Method of Name Discrimination Using Second Order Co-Occurrence Features

Ted Pedersen[1], Anagha Kulkarni[1], Roxana Angheluta[2],
Zornitsa Kozareva[3], and Thamar Solorio[4]

[1] University of Minnesota, Duluth, USA
[2] Katholieke Universiteit Leuven, Belgium
[3] University of Alicante, Spain
[4] University of Texas at El Paso, USA

Abstract. Previous work by Pedersen, Purandare and Kulkarni (2005) has resulted in an unsupervised method of name discrimination that represents the context in which an ambiguous name occurs using second order co–occurrence features. These contexts are then clustered in order to identify which are associated with different underlying named entities. It also extracts descriptive and discriminating bigrams from each of the discovered clusters in order to serve as identifying labels. These methods have been shown to perform well with English text, although we believe them to be language independent since they rely on lexical features and use no syntactic features or external knowledge sources. In this paper we apply this methodology in exactly the same way to Bulgarian, English, Romanian, and Spanish corpora. We find that it attains discrimination accuracy that is consistently well above that of a majority classifier, thus providing support for the hypothesis that the method is language independent.

1 Introduction

Purandare and Pedersen (e.g., [9], [10]) previously developed an unsupervised method of word sense discrimination that has also been applied to name discrimination by Pedersen, Purandare, and Kulkarni [8]. This method is characterized by a reliance on lexical features, and avoids the use of syntactic or other language dependent information. This is by design, since the method is intended to port easily and effectively to a range of languages. However, all previous results with this method have been reported for English only.

In this paper, we evaluate the hypothesis that this method of name discrimination is language independent by applying it to name discrimination problems in Bulgarian, Romanian, and Spanish, as well as in English.

Ambiguity in names of people, places and organizations is an increasingly common problem as online sources of information grow in size and coverage. For example, Web searches for names frequently locate different entities that share

the same name, and this can lead to considerable confusion or misunderstanding on the part of users.

This method assumes that Named Entity Recognition (NER) has already been performed on the text. Thus, our goal is not to identify named entities in text, but rather to disambiguate among those that have already been identified and determine the number of underlying entities that might share the same name.

This paper continues with an overview of our method of clustering similar contexts in order to perform name discrimination. Then we describe the experimental data for each of the four languages included in this study in some detail. We go on to present our experimental results, focusing on the overall accuracy of the automatic discrimination, and giving examples of the labels that are created for clusters. We close with a discussion of related work and some brief thoughts on future work.

2 Discrimination by Clustering Similar Contexts

The method of clustering similar contexts developed by Purandare and Pedersen is well described elsewhere (e.g., [9], [10]) and is implemented in the freely available SenseClusters package[1].

In this paper we employ one variation of their general approach, which results in a second order co-occurrence representation of the contexts to be clustered. We begin with a collection of contexts to be clustered. In general a context can be an unit of text from a few words to a paragraph or entire document. In these experiments, each context contains one or two sentences that contain a single occurrence of an ambiguous name.

If there are a small number of such contexts to cluster, it might be necessary to select the features to represent these contexts from a separate corpus (assuming it is relevant to the contexts to be clustered). However, in this case there are a sufficient number of contexts to cluster such that features can be identified within that data. Thus, in these experiments we say that the test or evaluation data is the same data as the feature selection data. These methods are completely unsupervised and the true senses of the ambiguous name are not used in the feature selection phase or at any stage of the method apart from the evaluation phase. Thus even if one uses the test data as the feature selection data this does not unfairly influence our results as it would for supervised methods.

We identify bigram features from the contexts to be clustered using the log–likelihood ratio. We define bigrams to be two word sequences where no intervening words are allowed. We conducted our experiments both with and without a stop–list, which is a list of closed–class words such as articles, conjunctions, prepositions, etc. When using stop–lists, any bigram made up of one or two stop words is rejected as a feature.

---

[1] http://senseclusters.sourceforge.net

In addition, any bigram that occurs only one time was rejected as a feature, as would be any bigram that has a log–likelihood ratio score less than 3.841. Bigrams with values under this threshold have a 95% chance of being independent of each other, that is they are occurring together as if by chance.

The bigram features are represented as a co–occurrence matrix, where the rows represent the first word in the bigram, and the columns represent the second word. The cell values are the corresponding log-likelihood ratios. Singular Value Decomposition (SVD) is performed on this matrix, reducing it down to 10% of the original number of columns, or to 300 dimensions, whichever is smaller. Each row in the resulting matrix is viewed as a co–occurrence vector for the word associated with that row.

We represent the contexts to be clustered using second order context vectors. These vectors are created by considering a *test scope* of five words to the left and five words to the right of the ambiguous target word. The words found in this window for which we have a co–occurrence vector are replaced by their vector. Then the context is represented by averaging together all the vectors found for the words in the test scope.

The resulting contexts were then clustered using the k-means clustering algorithm (referred to as direct clustering in CLUTO[2], which is the package SenseClusters uses for clustering). We used the I2 criterion function for clustering, and we must specify the number of clusters we wish to find prior to clustering. The I2 criterion function finds the clustering solution that minimizes the distance of the members of a cluster to the centroid of its cluster.

In the experiments in this paper we know what the "correct" clustering should be, since we will create ambiguous names by conflating together relatively unambiguous names and replacing each occurrence of each name with the newly ambiguous conflated form. Then, the effectiveness of the clustering can be evaluated by measuring how well the discovered clusters have separated the entities associated with the name we have conflated together. We report results in terms of accuracy, that is what percentage of the contexts are correctly clustered.

Finally, cluster labels are created for each cluster by identifying the top ten descriptive and discriminating bigrams according to the log–likelihood ratio found in the contexts of each cluster. These bigrams are found by treating the contexts in each cluster as a corpus, and applying the measure in exactly the same way as we did during feature identification. However, for labeling, we allow up to three intervening words between the words that make up the bigram. The descriptive bigrams are the top ten bigrams found in the contexts associated with a cluster, and the discriminating bigrams are the top ten bigrams that are unique in the contexts associated with a cluster. Thus, it's possible that the descriptive and discriminating labels will overlap.

---

[2] http://www-users.cs.umn.edu/~karypis/cluto/

3 Second Order Co–Occurrence Features

At the heart of this method lies the idea of a second order co–occurrence vector. In general, a second order co–occurrence exists between two words that do not occur together, but both tend to occur with some other word. For example, *fire* and *garden* might not occur together often, but both may occur frequently with *hose*, as in *fire hose* and *garden hose*. Thus, there is an indirect relationship between *fire* and *garden* through *hose*. This can be thought of as a friend of a friend relation.

Our method for creating second order features was originally proposed by Schütze[11]. It does not directly search for second order co–occurrences in the contexts to be clustered (by creating a network of word associations, for example). Rather, they are identified as a by–product of the method used for representing the contexts to be clustered. Recall that a word by word co–occurrence matrix is created from the feature selection data. This is a matrix containing information about first order co–occurrences, that is showing which words occur together. Each word in a context to be clustered is represented by a vector from this word by word matrix (if one exists for that word), which indicates the first order co–occurrences of that word.

Once collected, all of the available word vectors associated with a context are averaged together to form a representation of that context. Remember that the context contains an occurrence of the ambiguous name, whose underlying identity is what we seek to base our clustering upon. Thus, the name to be disambiguated is represented not by the words that it occurs with, but rather by the average of the first order vectors of the words that co-occur with the target name. Thus, the name to be disambiguated is represented by second order co–occurrences.

We believe second order features are a suitable representation for this problem, since they allow us to find more matching features when confronted with relatively sparse or noisy data. While our data occurs in fairly large quantities, it is from newspaper corpora and as such can be somewhat unfocused or rapidly changing.

4 Experimental Data

In order to evaluate the language independence of our method, we utilized four languages in our experiments: Bulgarian, English, Spanish, and Romanian. We have at least one native speaker of each language among the authors of the paper.

We located large news corpora for each of the four languages, and then identified named entities automatically, or based on our own knowledge of current events and regional history. In order to facilitate evaluation, we created ambiguities in the data by conflating together names that are largely unambiguous. For example, we took all occurrences of *Bill Clinton* and all occurrences of *Tony*

Blair and made their names ambiguous by replacing them with *Bill Clinton-Tony Blair*[3].

These conflated names appear ambiguous to our method, but we of course know their true underlying identity (pre–conflation) which will allow for automatic evaluation. The methodology and motivation behind creating conflated names are identical to pseudo–words as used in the word sense disambiguation literature. One known drawback of pseudo–words arises when the component words are randomly selected. In such a case, it is very likely that the two senses represented will be quite distinct [2]. However, our formulation is similar to that of Nakov and Hearst [7], who suggest creating pseudo words of words that are individually unambiguous, and yet still related in some way.

For each language we created five sets of conflations for use in our experiments. Two sets contained the names of people, two contained country or city names, and then one set included organization names. Thus we are making distinctions between names that are of the same general class, making these less obvious distinctions than those between a city and a person, for example. We did not use the same names for all of the languages, since some of the names were specific to a particular language or region and would not appear in sufficient quantity for experimenting in all languages. However, the fact that the words share general categories makes the results somewhat comparable.

For each language also we found or manually constructed a stop–list of commonly used words, consisting mostly of function words such as articles, conjunctions, and so forth. The stop–lists were of comparable size, except for Bulgarian which was somewhat larger with 806 stop words. English had 426, Romanian 438, and Spanish 499. In fact, the stop–lists are not derived in a language independent way for these experiments, and represent the only language dependent part of the process. However, we believe that it will be possible to develop methods that derive stop–lists automatically. This remains an important area of future work.

Below we describe the names used in our experiments, and the corpora from which they were derived. We provide a brief description of each named entity. Note that the distribution of names prior to conflation is shown in Table 1.

4.1 Bulgarian

The Bulgarian experiments relied on the Sega2002 news corpus, which was originally prepared for the CLEF[4] competition. This is a corpus of news articles from the Newspaper Sega[5], which is based in Sofia, Bulgaria.

The version of the corpus used in our experiments was created with the help of the CLaRK system[6]. Initially individual articles were found in different XML files

---

[3] The actual conflation of the data was done with version 0.14 of the freely available nameconflate program (http://www.d.umn.edu/~kulka020/kanaghaName.html).

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

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

[6] http://bultreebank.org/clark/

depending on the year, month, and day of their publication. We merged these into a single file and only utilized the content between the text tags. The sentences that contained the names to be used in the experiments were extracted, and the Cyrillic characters were transliterated. Most Cyrillic characters are mapped one to one to the Latin alphabet, however several Cyrillic characters had to be represented by combination of two Latin symbols as the transliteration was phonetically based.

The Bulgarian stop–list was taken from the resources distributed with the HPSG-based Syntactic Treebank of Bulgarian[7].

Countries Germaniya (Germany), Franciya (France), and Rusiya (Russia) are major European countries. Their occurrences were conflated into a single three way ambiguous name Fr-Ge-Ru.

Organizations The organization names in this experiment are the abbreviations of the two leading political parties in Bulgaria. BSP (Balgarska Socialisticeska Partija, or Bulgarian Socialist Party) is the left leaning party and the successor to the Bulgarian Communist Party. It was formed in 1990 in post-communist Bulgaria. SDS (Saíuz na demokratichnite sili, or The Union of Democratic Forces) is the right leaning political party. It was formed at the end of 1989 as a union of non–governmental organizations and reinvigorated old parties who had historically opposed the Communist government. These two names were conflated into a single ambiguity, BSP-SDS.

Cities Varna and Burgas are the largest cities on the Bulgarian Black Sea Coast, and are the third and fourth largest cities overall in Bulgaria. Their names were combined into a single ambiguity, Va-Bu.

People Ivan Kostov was Prime Minister of Bulgaria from May 1997 to June 2001 and leader of the Union of Democratic Forces (SDS, see above) between December 1994 and June 2001. Presently he is leader of the political party he formed, Democrats for a Strong Bulgaria. Petar Stoyanov was the President of Bulgaria from 1998 until 2002. He is now chairman of the Union of Democratic Forces (SDS, see above). Georgi Parvanov has been the President of Bulgaria since January 22, 2002. He is member of the Bulgarian Communist Party. The three names above were conflated to form the ambiguous name PS-IK-GP.

Nadejda Mihaylova is politician from the Democratic party and was Minister of Exterior from 1997 to 2001. Nikolay Vasilev is a politician from the National Movement Simeon II party. He was Vice-Premier and Minister of Economics during 2001, and Vice-Premier and Minister of Transport and Communications during 2003. Simeon Sakskoburggotski was the last King of Bulgaria, and was Prime Minister of Bulgaria from 2001 until August 2005. These three names were conflated to form the ambiguous name NM-NV-SS.

---

[7] http://www.bultreebank.org/Resources.html

4.2 Romanian

The Romanian data was taken from the 2004 archives of the newspaper Adevarul (The Truth)[8]. This is a daily newspaper that is among the most popular in Romania. Named entities of interest were extracted via the grep command, and then any remaining html tags were removed. While Romanian typically contains diacritical markings, Adevarul does not publish their text with diacritics, so it was not necessary to account for them in processing.

We initially used a stop–list created by Rada Mihalcea[9], but observed that it was somewhat smaller than the stop–lists we were using for the other languages. It had approximately 250 entries, whereas the English and Spanish stop–lists had more than 400 entries, and Bulgarian approximately 800. Thus, we augmented the stop–list to make it more comparable with the other languages, so that the version we used in our experiments has 438 stop words.

The original Mihalcea stop–list followed pre-Revolution spelling conventions. For example, prior to 1989 verbs like *a minca* (to eat) were spelled *mînca* (*minca* after removing diacritics) while now they are spelled *mânca* (*manca* after removing diacritics). Another example is the verb *to be* which, for first person, was spelled *sînt* (I am) while now it is spelled *sunt*. The words following post-Revolution conventions have been added to the list. Another source of new words was an online Romanian dictionary[10], which offered all the inflected forms for pronouns. As a general remark, since Romanian is a language with a rich morphology, when adding a new word to the stop–list generally all the inflected forms have been added as well. Finally, the list was enriched also by translating words from the English stop–list, when appropriate.

Organizations Partidul Democrat (PD) is the Romanian Democratic Party. For the 2004 elections they joined forces with the National Liberal Party to create the Justice and Truth (Dreptate si Adevar) political alliance, whose main purpose was to compete against PSD. They were successful in this election, and now hold power in Romania. The Partidul Social Democrat (PSD) is currently the main opposition party in Romania. These two names were conflated into the ambiguous name PD-PSD.

People Traian Basescu is the current president of Romania, elected in 2004. His principal rival for the presidency was Adrian Nastase. Between 2000 and 2004 Basescu was the mayor of Bucharest. His political party is Partidul Democrat (PD, see above). Adrian Nastase is currently the President of Chamber of Deputies. In 2004 he competed for the presidential elections but he was defeated by Traian Basescu. He was Prime Minister between 2000 and 2004. He is a member of the Partidul Social Democrat (PSD) (see above). These names were conflated to create a two way ambiguity, TB-AN.

---

[8] http://www.adevarulonline.ro/arhiva
[9] http://nlp.cs.nyu.edu/GMA_files/resources/romanian.stoplist
[10] http://dexonline.ro/

Ion Iliescu is the former Romanian president. He was president for 11 years, from 1990 to 1996 and from 2000 to 2004. Currently he is a senator for the PSD. This name was added to the two above to create a three way ambiguity, TB-II-AN.

Cities Bucuresti (Bucarest) is the Romanian capital. It is the largest city in Romania, located in the southeast of the country. Brasov is a popular tourist destination in central Romania, located in the Carpathian Mountains of Transylvania. These two names were conflated into a single ambiguity Br-Bu.

Countries The country names included in the Romanian experiment include Romania, SUA (Statele Unite ale Americii, or the United States), and Franta (France). Their names were conflated into a single ambiguity, Fr-SUA-Ro.

4.3 English

The source of the English data was the English GigaWord Corpus, available from the Linguistic Data Consortium. In total this contains 1.7 billion words from four different news services. Our data was selected from either the 900 million word New York Times (nyt) portion or the 170 million word Agence France Presse English Service (afe) portion of the corpus. This text comes from the period 1994 through 2002.

Organizations Microsoft is the world's largest software company. It was founded in 1975 by Bill Gates and Paul Allen. IBM is a large computer hardware and software company that has existed since 1888. These names were conflated into IBM-Mi.

Locations There were three countries and one state included in these experiments. Mexico is the largest Spanish-speaking country in the world. It is located in North America, directly south of the United States. Uganda is a country in East Africa. While it is landlocked, it has access to Lake Victoria, the largest lake in Africa. These two country names were conflated into Me-Ug.

India is a South Asian country which is the second most populous in the world. California is the most populous state in the United States. It is on the west coast. Peru is a Spanish speaking country in western South America. These four names were conflated into Me-In-Ca-Pe.

People Tony Blair is the current Prime Minister of England. He has held this office since 1997. He is the leader of the Labour Party. Bill Clinton was the 42nd President of the United States, and was in office from 1993 to 2001. Prior to serving as President, he was the Governor of Arkansas. He is a member of the Democratic Party. These two names were conflated into BC-TB.

Ehud Barak was the 10th Prime Minister of Israel, serving from 1999 to 2001. He was the leader of the Labor Party. This name was added to the two above to create the three way ambiguity, BC-TB-EB.

4.4 Spanish

The Spanish corpora comes from the Spanish news agency EFE from the year 1994. It contains a total of 215,738 documents. This collection was used in the Question Answering Track at CLEF-2003, and also for CLEF-2005.

A Named Entity Recognizer was used, and then the frequencies of entities was manually examined to determine the list of candidates for the experiment. The stop–list for Spanish was the same as used in the CLEF-2005 competition.[11].

People Yaser Arafat was the Chairman of the Palestine Liberation Organization from 1969 until his death in 2004. Bill Clinton is a former US president, as mentioned above. These two names were conflated to the ambiguous form YA-BC.

Juan Pablo II (John Paul II) was pope of the Roman Catholic Church from 1987 until his death in 2005. Boris Yeltsin was the President of Russia from 1991 to 1999. These were conflated to JP-BY.

Organizations OTAN is the Spanish abbreviation for NATO, the North Atlantic Treaty Organization. This is an alliance between the United States, Canada and many European nations. EZLN is the Ejército Zapatista de Liberación Nacional, known in English as the Zapatista Army of National Liberation. It is based in Chiapas, Mexico and seeks to make revolutionary changes to Mexican society. These two names were conflated to form OTAN-EZLN.

Cities Nueva York (New York) and Washington are major cities in the United States. Washington may also refer to a state on the West coast of the USA, so there is some ambiguity. These were conflated to form NY-Wa.

Brasil (Brazil) is the largest country in South America, both in terms of land mass and population. This was added to the names above to form the conflation NY-Br-Wa.

5 Experimental Results and Discussion

Our experimental results are summarized in Table 1. The conflated names are shown in the first column, and then the distribution of the instances for each underlying entity in the name are shown. For example, we can see that the conflated Bulgarian name Va-Bu occurred 2,501 times, and that 1,240 of these were the underlying entity Varna, and 1,261 were for Burgas.

Please note that the names are organized for each language such that the first two entries are for people, the third entry for organizations, and the last two are for locations.

In the third column the percentage of the instances that belong to the most frequent underlying entity are shown. This value is associated with a simple baseline clustering method that would simply assign all of the contexts to one

---

[11] http://www.unine.ch/info/clef

Name	Distribution	Majority	No Stop–list		With Stop–list		Diff.
Bulgarian:							
PS-IK-GP	318+524+811=1653	49.06%	40.53%	670	**58.68%**	970	+9.62
NM-NV-SS	645+849+976=2470	39.51%	35.79%	884	**59.39%**	1467	+19.88
BSP-SDS	2921+4680=7601	61.57%	51.97%	3950	57.31%	4356	-4.26
Fr-Ge-Ru	1726+2095+2645=6466	40.91%	39.07%	2526	41.60%	2690	+0.69
Va-Bu	1240+1261=2501	50.42%	**66.09%**	1653	50.38%	1260	+15.71
English:							
BC-TB	1900+1900=3800	50.00%	**80.89%**	3074	**80.95%**	3076	+30.95
BC-TB-EB	1900+1900+1900=5700	33.33%	**46.68%**	2661	**47.93%**	2732	+14.60
IBM-Mi	2406+3401=5807	58.57%	50.59%	2938	**63.70%**	3699	+5.13
Me-Ug	1256+1256=2512	50.00%	50.76%	1275	**59.16%**	1486	+9.16
Me-In-Ca-Pe	1500+1500+1500+1500=6000	25.00%	**28.75%**	1725	**28.78%**	1727	+3.78
Romanian:							
TB-AN	1804+1932=3736	51.34%	50.59%	1890	51.34%	1918	+0.00
TB-II-AN	1948+1966+2301=6215	37.02%	34.16%	2123	**39.31%**	2443	+2.29
PD-PSD	2037+3264=5301	61.57%	52.08%	2761	**77.70%**	4119	+16.13
Br-Bu	2310+2559=4869	52.56%	51.22%	2494	**63.67%**	3100	+11.11
Fr-SUA-Ro	1370+2396+3890=7656	50.81%	40.73%	3118	52.66%	4032	+1.85
Spanish:							
YA-BC	1004+2340=3344	69.98%	50.24%	1680	**77.72%**	2599	+7.74
JP-BY	1447-1450=2897	50.05%	63.62%	1843	**87.75%**	2897	+37.70
OTAN-EZLN	1093+1093=2186	50.00%	50.09%	1095	**69.81%**	1526	+19.81
NY-Wa	1517+2418=3935	61.45%	54.69%	2152	54.66%	2151	-6.76
NY-Br-Wa	1517+1748+2418=5863	42.55%	39.24%	2230	42.88%	2437	+0.33

cluster. Then columns 4 and 5 show the accuracy associated with the clustering without and with a stop–list. The number of contexts that are clustered correctly are shown next to the accuracy percentage. Finally, the last column shows the difference between the best result obtained for a name with the majority percentage.

Generally we can observe that nearly all of the experiments show a positive increase from the majority classifier. In nearly all cases the best results are shown when using a stop–list. Of the 20 conflated names, 4 show a significant increase above the majority class without using a stop–list, and 13 show an significant increase beyond the majority class with a stop–list.

In addition, in nearly all cases the use of a stop–list either results in better or equally good accuracy as when not using a stop–list. The only exception to this is the Bulgarian name Va-Bu, which has two underlying entities, the cities of Varna and Burgas. This is the only case where the use of a stop–list has actually hurt performance rather badly. However, we also note that the location names in general seem to show relatively little improvement with stop–lists in at least one of the two cases for all of the languages.

We theorize that location names can be used in a wide range of contexts, and that it is harder to find discriminating features for them. However, locations also have many unique names associated with them, so it is still unclear to us why the location names often pose the most significant challenges for this approach.

6   Cluster Label Examples

In addition to grouping the contexts into clusters that reflect the underlying or true entities, we generate a label for each cluster based on its content. This is intended to act as a simple identifier for the underlying entity.

The top ten bigrams found in the contexts in a cluster that do not include stop words and have a log-likelihood score above 3.841 are chosen as the descriptive labels, regardless of how many other clusters they may occur in. The discriminating labels are the top ten bigrams that are unique to a cluster, according to these same criteria. In the cluster labels we allow the words that form a bigram to be separated by up to three intervening words.

As yet we do not have a reliable means of evaluating these labels, so we simply show examples of the labels found for each language in Table 2. For each language we show a two sense distinction, where the true underlying entity for a cluster is on the left, and the automatically generated labels are on the right. The descriptive labels are shown in normal text, and labels that are both discriminating and descriptive are in bold. Note that descriptive labels may be shared by the two clusters, and can be thought of as providing some indication of the general topic or subject that pertains to both clusters. The discriminating labels are meant to distinguish between the different clusters.

In these examples there is no discriminating label that is not a descriptive label as well. This simply indicates that all of the discriminating labels occurred in the top ten bigrams overall.

For Bulgarian and Romanian, we show the cases where two political parties are discriminated. The labels consist mainly of names, and in general these names are commonly understood to be associated with the party mentioned.

In Bulgarian the BSP cluster shows discriminating labels that include *liderat Sergey* and *liderat Stanishev*, which is quite reasonable since *liderat* means *leader*, and *Sergey Stanishev* is the leader of the Socialist Party in Bulgaria (BSP). Also note that he appears as a descriptive label for SDS. This can be understood by pointing out that the leader of an opposing party could well be mentioned in contexts that are about the SDS. It is encouraging to note that references to him as *leader* were unique to the BSP cluster.

In Romanian, the PSD cluster includes *Ion Iliescu* and *Cozmin Gusa* as discriminating features, both who are members of the PSD. The PSD cluster also has *partidul guvernamant* as a discriminating feature, which means *government party*, which describes the PSD in 2004. The PD cluster includes discriminating labels *Popescu Tariceanu*, *Theodor Stolojan*, and *Calin Tariceanu*, who are all members of the Liberal Party, which formed an alliance with the PD. And in fact that alliance has been included as a discriminating label via *Alianta PNL*.

Table 2. Cluster Label Examples

True Entity	Created Labels
Bulgarian (2 political parties)	
BSP	**Nikolay Mladenov, liderat Sergey, Visshiya savet, liderat Stanishev,** vot nedoverie, Ekaterina Mihaylova, Sergey Stanishev, G n, Ivan Kostov, Nadejda Mihaylova
SDS	**mestnite izbori, d r, Rakovski 134, politicheska sila,** vot nedoverie, Ekaterina Mihaylova, Sergey Stanishev, G n, Ivan Kostov, Nadejda Mihaylova
English (2 companies)	
IBM	**5 8, BW GEN, 3 4, interest rates, Texas Instruments, Hewlett Packard, 30 Treasury, 7 8,** Wall Street, billion dollars
MICROSOFT	**vice president, million dollars, Windows 95, operating system United States, Bill Gates, Justice Department, personal computers,** Wall Street
Romanian (2 political parties)	
PD	**Popescu Tariceanu, Theodor Stolojan, Alianta PNL, Calin Tariceanu,** Camera Deputatilor, PNL PD, Adrian Nastase, Traian Basescu
PSD	**Camera Uniunea, Deputatilor Uniunea, partidul guvernamant, Cozmin Gusa, Ion Iliescu, Emil Boc,** Camera Deputatilor, PNL PD, Adrian Nastase, Traian Basescu, Emil Boc
Spanish (2 leaders)	
BILL CLINTON	**presidente estadounidense, EFE presidente, presidente OLP, Casa Blanca, Washington EFE, presidente Unidos**
YASER ARAFAT	**Exteriores Peres, ministro israeli, Palestina OLP, Gaza Jerico, Hafez Asad, Isaac Rabin, proceso paz, Asuntos Exteriores**

Note that the full name of the alliance is *Alianta PNL PD*, but since we rely on bigrams this has been split into two (where *PNL PD* is included as a descriptive label of PD).

In English we show the labels for the clusters associated with IBM and Microsoft. We note that these labels are somewhat noisier than those of the political parties. For example, there are a number of unusual looking pairs of numbers in the IBM cluster. However, these are the result of a tokenization scheme that simply removed non-alphanumeric characters (e.g., so *3/8* become *3 8*). These fractions refer to movements in the stock price. The companies *Texas Instruments* and *Hewlett Packard* are shown as discriminating labels for IBM, and may reflect the fact that these companies are often mentioned together when discussing stock market activity. The inclusion of *Wall Street* and *billion dollars* as descriptive labels for IBM is due to the financial focus of much of the news reporting, but does not offer much in terms of unique information about IBM.

The Microsoft cluster has a discriminating label *Bill Gates*, who is the co–founder of the company. It also includes *Justice Department* as a discriminating label, which is appropriate given the great attention paid to the legal case against Microsoft. The discriminating labels *Windows 95*, *operating system*, and *personal computers* are certainly useful in identifying Microsoft, whereas those for *vice president* and *million dollars* are less so.

In the Spanish data, all of the labels shown are both descriptive and discriminating, meaning that the top ten bigrams in each cluster were unique to that cluster. The labels for Bill Clinton include *presidente estadounidense*, which translates as *President of the United States*, and *Casa Blanca*, which is the White House. It also has a certain amount of noise, for example various labels that mention EFE, which is a Spanish news agency and in fact the source of this corpora. We believe that this is due to the presence of datelines in the contexts, as in *Washington, Jan 2 (EFE) - President Clinton said ….*

The labels for Yaser Arafat include several that are quite discriminating, including *processo paz* (*peace process*), and *Palestina OLP*, which refers to the Palestinian Liberation Organization (OLP in Spanish). However, the cluster for Bill Clinton also includes *presidente OLP*, due to his frequent meetings with Arafat during this time. *Hafez Asad* was the president of Syria, and *Isaac Rabin* was the Prime Minister of Israel (known as Yitzhak in English).

In general we can see that these labels provide relevant and useful information about the underlying entities, but that they are somewhat noisy and perhaps not obvious indicators of that entity. Please note that the descriptive labels are not intended to uniquely describe the cluster, but rather to give an overall gist of what the cluster is about, while the discriminating labels are those that are meant to provide the unique information about an underlying identity.

7   Related Work

Bagga and Baldwin [1] propose a method for resolving cross document references (such as recognizing that John Smith and Mr. Smith refer to the same person) based on creating first order context vectors that represent each instance in which an ambiguous name occurs. Each vector contains exactly the words that occur within a 55 word window around the ambiguous name, and the similarity among names is measured using the cosine measure. In order to evaluate their approach, they created the *John Smith* corpus, which consists of 197 articles from the New York Times that mention 35 different *John Smiths*.

Gooi and Allan [4] present a comparison of Bagga and Baldwin's approach to two variations of their own. They used the *John Smith* Corpus, and created their own corpus which is called the *Person-X* corpus. Since it is rather difficult to obtain large samples of data where the actual identity of a truly ambiguous name is known, the *Person-X* corpus consists of pseudo-names that are ambiguous. These are created by disguising known names as *Person-X*, thereby introducing ambiguities. There are 34,404 mentions of *Person-X*, which refer to 14,767 distinct underlying entitles. Gooi and Allan re–implement Bagga and Baldwin's context vector approach, and compare it to another context vector approach that groups vectors together using agglomerative clustering. They also group instances together based on the Kullback–Liebler Divergence. Their conclusion is that the agglomerative clustering technique works particularly well.

Mann and Yarowsky [6] have proposed an approach for disambiguating personal names using a Web based unsupervised clustering technique. They rely

on a rich feature space of biographic facts, such as date or place of birth, occupation, relatives, collegiate information, etc. A seed fact pair (e.g., Mozart, 1776), is queried on the Web and the sentences returned as search results are used to generate the patterns which are than used to extract the biographical information from the data. Once these features are extracted clustering follows. Each instance of an ambiguous name is assigned a vector of extracted features, and at each stage of cluster the two most similar vectors are merged together to produce a new cluster. This step is repeated until all the references to be disambiguated are clustered.

Name disambiguation is also a problem in the medical domain. For example, Hatzivassiloglou, et. al. [5] point out that genes and proteins often share the same name, and that it's important to be able to identify which is which. They employ a number of well known word sense disambiguation techniques and achieve excellent results. Ginter, et. al. [3] develop an algorithm for disambiguation of protein names based on weighted features vectors derived from surface lexical features and achieve equally good results.

8   Future Work

There are two language dependent aspects to this method. The first is that it does assume that the words in the language have been segmented. In the case of the languages used in this study, we have simply assumed words to be alphabetic strings that are white space separated. However, in some languages segmentation is a more difficult issue, and that would need to be resolved before this method was applied.

Second, we have utilized pre-existing or manually derived stop–lists, which introduces a language dependence on our method. We are confident that we can develop a language independent method of finding stop words in the corpora we are clustering. Some variant of term frequency/inverse document frequency (TF/IDF) might be appropriate, or we could simply identify those words that occur in a majority of all contexts and consider those as stop words.

9   Conclusions

The experiments and results in this paper show that our hypothesis that these methods are language independent has some validity. Results well in excess of the majority class baseline are obtained for four different languages using exactly the same methodology. The fact that these methods are completely unsupervised and yet they could be successfully applied to the discrimination problem from different domains like politics, geographical locations, and organizations also suggests that the methods are also domain–independent.

10 Acknowledgments

References

1. A. Bagga and B. Baldwin. Entity–based cross–document co–referencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics*, pages 79–85. Association for Computational Linguistics, 1998.

2. T. Gaustad. Statistical corpus-based word sense disambiguation: Pseudowords vs. real ambiguous words. In *Proceedings of the Student Research Workshop at ACL-2001*, pages 61–66, Toulouse, France, 2001.

3. F. Ginter, J. Boberg, J. Jrvine, and T. Salakoski. New techniques for disambiguation in natural language and their application to biological text. *Journal of Machine Learning Research*, 5:605–621, June 2004.

4. C. H. Gooi and J. Allan. Cross-document coreference on a large scale corpus. In S. Dumais, D. Marcu, and S. Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 9–16, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

5. V. Hatzivassiloglou, P. Duboue, and A. Rzhetsky. Disambiguating proteins, genes, and rna in text: A machine learning approach. In *Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology*, Tivoli Gardens, Denmark, July 2001.

6. G. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 33–40. Edmonton, Canada, 2003.

7. P. Nakov and M. Hearst. 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 2003.

8. T. Pedersen, A. Purandare, and A. Kulkarni. 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 2005.

9. A. Purandare. Discriminating among word senses using McQuitty's similarity analysis. In *Proceedings of the Student Research Workshop at HLT-NAACL 2003*, pages 19–24, Edmonton, Alberta, Canada, May 27 - June 1 2003.

10. A. Purandare and T. Pedersen. 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, 2004.

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