# A Language Independent Approach for Name Categorization and Discrimination


**Zornitsa Kozareva**
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
zkozareva@dlsi.ua.es

**Sonia Vázquez**
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
svazquez@dlsi.ua.es

**Andrés Montoyo**
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
montoyo@dlsi.ua.es



## Abstract

We present a language independent approach for fine-grained categorization and discrimination of names on the basis of text semantic similarity information. The experiments are conducted for languages from the Romance (Spanish) and Slavonic (Bulgarian) language groups. Despite the fact that these languages have specific characteristics as word-order and grammar, the obtained results are encouraging and show that our name entity method is scalable not only to different categories, but also to different languages. In an exhaustive experimental evaluation, we have demonstrated that our approach yields better results compared to a baseline system.


## 1 Introduction

### 1.1 Background

Named Entity (NE) recognition concerns the detection and classification of names into a set of categories. Presently, most of the successful NE approaches employ machine learning techniques and handle simply the person, organization, location and miscellaneous categories. However, the need of the current Natural Language Applications impedes specialized NE extractors which can help for instance an information retrieval system to determine that a query about "Jim Henriques guitars" is related to the person "Jim Henriques" with the semantic category musician, and not "Jim Henriques" the composer. Such classification can aid the system to rank or return relevant answers in a more accurate and appropriate way.

So far, the state-of-art NE recognizers identify that "Jim Henriques" is a person, but do not subcategorize it. There are numerous drawbacks related to the fine-grained NE issue. First, the systems need hand annotated data which are not available for multiple categories, because their creation is time-consuming, requires supervision by experts, a predefined fine-grained hierarchical structure or ontology. Second, there is a significant lack of freely available or developed resources for languages other than English, and especially for the Eastern European ones.

The World Wide Web is a vast, multilingual source of unstructured information which we consult daily in our native language to understand what the weather in our city is or how our favourite soccer team performed. Therefore, the need of multilingual and specialized NE extractors remains and we have to focus on the development of language independent approaches.

Together with the specialized NE categorization, we face the problem of name ambiguity which is related to queries for different people, locations or companies that share the same name. For instance, Cambridge is a city in the United Kingdom, but also in the United States of America. ACL refers to "The Association of Computational Linguistics", "The Association of Christian Librarians" or to the "Automotive Components Limited". Googling the name "Boyan Bonev" returns thousands of documents where some are related to a member of a robot vision group in Alicante, a teacher at the School



*Balto-Slavonic Natural Language Processing 2007, June 29, 2007*, pages 19–26,
Prague, June 2007. 

of Biomedical Science, a Bulgarian schoolboy that participated in computer science competition among others. So far, we have to open the documents one by one, skim the text and decide to which "Boyan Bonev" the documents are related to. However, if we resolve the name disambiguation issue, this can lead to an automatic clustering of web pages talking about the same individual, location or ogranization.

### 1.2 Related Work

Previously, (Pedersen et al., 2005) tackled the name discrimination task by developing a language independent approach based on the context in which the ambiguous name occurred. They construct second order co-occurrence features according to which the entities are clustered and associated to different underlying names. The performance of this method ranges from 51% to 73% depending on the pair of named entities that have to be disambiguated. Similar approach was developed by (Bagga and Baldwin, 1998), who created first order context vectors that represent the instance in which the ambiguous name occurs. Their approach is evaluated on 35 different mentions of John Smith, and the f-score is 84%.

For fine-grained person NE categorization, (Fleischman and Hovy, 2002) carried out a supervised learning for which they deduced features from the local context in which the entity resides, as well as semantic information derived from the topic signatures and WordNet. According to their results, to improve the 70% coverage for person name categorization, more sophisticated features are needed, together with a more solid data generation procedure. (Tanev and Magnini, 2006) classified geographic location and person names into several subclasses. They use syntactic information and observed how often a syntactic pattern co-occurs with certain member of a given class. Their method reaches 65% accuracy. (Pasca, 2004) presented a lightly supervised lexico-syntactic method for named entity categorization which reaches 76% when evaluated with unstructured text of Web documents.

(Mann, 2002) populated a fine-grained proper noun ontology using common noun patterns and following the hierarchy of WordNet. They studied the influence of the newly generated person ontology in a Question Answering system. According to the obtained results, the precision of the ontology is high, but still suffers in coverage. A similar approach for the population of the CyC Knowledge Base (KB) was presented in (Shah et al., 2006). They used information from the Web and other electronically available text corpora to gather facts about particular named entities, to validate and finally to add them to the CyC KB.

In this paper, we present a new text semantic similarity approach for fine-grained person name categorization and discrimination which is similar to those of (Pedersen et al., 2005) and (Bagga and Baldwin, 1998), but instead of simple word co-occurrences, we consider the whole text segment and relate the deduced semantic information of Latent Semantic Analysis (LSA) to trace the text cohesion between thousands of sentences containing named entities which belong to different fine-grained categories or individuals. Our method is based on the word sense discrimination hypothesis of Miller and Charles (1991) according to which words with similar meaning are used in similar context, hence in our approach we assume that the same person or the same fine-grained person category appears in the similar context.

## 2 NE categorization and discrimination with Latent Semantic Analysis

LSA has been applied successfully in many areas of Natural Language Processing such as Information Retrieval (Deerwester et al., 1990), Information Filtering (Dumais, 1995) , Word Sense Disambiguation (Shütze, 1998) among others. This is possible because LSA is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in discourse. It uses no humanly constructed dictionaries or knowledge bases, semantic networks, syntactic or morphological analyzers, because it takes only as input raw text which is parsed into words and is separated into meaningful passages. On the basis of this information, LSA extracts a list of semantically related word pairs or rank documents related to the same topic.

LSA represents explicitly terms and documents in a rich, highly dimensional space, allowing the underlying "latent", semantic relationships between terms and documents to be exploited. LSA relies on the constituent terms of a document to suggest the document's semantic content. However, the LSA model views the terms in a document as somewhat unreliable indicators of the concepts contained in the document. It assumes that the variability of word choice partially obscures the semantic structure of the document. By reducing the original dimensionality of the term-document space with Singular Value Decomposition to a matrix of 300 columns, the underlying, semantic relationships between documents are revealed, and much of the "noise" (differences in word usage, terms that do not help distinguish documents, etc.) is eliminated. LSA statistically analyzes the patterns of word usage across the entire document collection, placing documents with similar word usage patterns near to each other in the term-document space, and allowing semantically-related documents to be closer even though they may not share terms.

Taking into consideration these properties of LSA, we thought that instead of constructing the traditional term-document matrix, we can construct a term-sentence matrix with which we can find a set of sentences that are semantically related and talk about the same person. The rows of the term-sentence matrix correspond to the words of the sentence where the NE has to be categorized or discriminated (we call this sentence target sentence), while the columns correspond to the rest of the sentences with NEs. The cells of the matrix show the number of times a given word from the target sentence co-occurs in the rest of the sentences. When two columns of the term-sentence matrix are similar, this means that the two sentences contain similar words and are therefore likely to be semantically related. When two rows are similar, then the corresponding words occur in most of the same sentences and are likely to be semantically related.

In this way, we can obtain semantic evidence about the words which characterize a given person. For instance, a *football player* is related to words as *ball*, *match*, *soccer*, *goal*, and is seen in phrases such as "X *scores a goal*", "*Y is penalized*". Meanwhile, a *surgeon* is related to words as *hospital*, *patient*, *operation*, *surgery* and is seen in phrases such as "*X operates Y*", "*X transplants*". Evidently, the category football player can be distinguished easily from that of the surgeon, because both person names occur and relate semantically to different words.

Another advantage of LSA is its property of language independence, and the ability to link several flexions or declanations of the same term. This is especially useful for the balto-slavonic languages which have rich morphology. Once the term-sentence approach is developed, practically there is no restrain for LSA to be applied and extended to other languages. As our research focuses not only on the resolution of the NE categorization and discrimination problems as a whole, but also on the language independence issue, we considered the LSA's usage are very appropriate.

## 3 Development Data Set

For the development of our name discrimination and classification approach, we used the Spanish language. The corpora we worked with is the EFE94-95 Spanish news corpora, which were previously used in the CLEF competitions[1]. In order to identify the named entities in the corpora, we used a machine learning based named entity recognizer (Kozareva et al., 2007).

For the NE categorization and discrimination experiments, we used six different named entities, for which we assumed a-priory to belong to one of the two fine-grained NE categories `PERSON_SINGER` and `PERSON_PRESIDENT`. The president names are Bill Clinton, George Bush and Fidel Castro, and the singer names are Madonna, Julio Iglesias and Enrique Iglesias. We have selected these names for our experiment, because of their high frequency in the corpora and low level of ambiguity.

Once we have selected the names, we have collected a context of 10, 25, 50 and 100 words from the left and from the right of the NEs. This is done in order to study the influence of the context for the NE discrimination and categorization tasks, and especially how the context window affects LSA's performance. We should note that the context for the NEs is obtained from the text situated between the text tags. During the creation of the context window, we used only the words that belong to the document in which the NE is detected. This restriction is imposed, because if we use words from previous or following documents, this can influence and change

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

the domain and the topic in which the NE is seen. Therefore, NE examples for which the number of context words does not correspond to 10, 25, 50 or 100 are directly discarded.

From the compiled data, we have randomly selected different NE examples and we have created two data sets: one with 100 and another with 200 examples per NE. In the fine-grained classification, we have substituted the occurrence of the president and singer names with the obfuscated form `President_Singer`. While for the NE discrimination task, we have replaced the names with the `M_EI_JI_BC_GB_FC` label. The first label indicates that a given sentence can belong to the president or to the singer category, while the second label indicates that behind it can stand one of the six named entities. The NE categorization and discrimination experiments are carried out in a completely unsupervised way, meaning that we did not use the correct name and name category until evaluation.

# 4 Experimental Evaluation

## 4.1 Experimental Settings

As mentioned in Section 2, to establish the semantic similarity relation between a sentence with an obfuscated name and the rest of the sentences, we use LSA[2]. The output of LSA is a list of sentences that best matches the target sentence (e.g. the sentence with the name that has to be classified or discriminated) ordered by their semantic similarity score. Strongly similar sentences have values close to 1, and dissimilar sentences have values close to 0.

In order to group the most semantically similar sentences which we expect to refer to the same person or the same fine-grained category, we apply the graph-based clustering algorithm PoBOC (Cleuziou et al., 2004). We construct a new quadratic sentence-sentence similarity matrix where the rows stand for the sentence we want to classify, the columns stand for the sentences in the whole corpus and the values of the cells represent the semantic similarity scores derived from LSA.

On the basis of this information, PoBOC forms two clusters whose performance is evaluated in terms of precision, recall, f-score and accuracy which can be derived from Table 1.

[2] http://infomap-nlp.sourceforge.net/

| number of | Correct `PRESIDENT` | Correct `SINGER` |
|---|---|---|
| Assigned `PRESIDENT` | a | b |
| Assigned `SINGER` | c | d |

Table 1: Contingency table

We have used the same experimental setting for the name categorization and discrimination problems.

## 4.2 Spanish name categorization

In Table 2, we show the results for the Spanish fine-grained categorization. The detailed results are for the context window of 50 words with 100 and 200 examples. All runs, outperform a simple baseline system which returns for half of the examples the fine-grained category `PRESIDENT` and for the rest `SINGER`. This 50% baseline performance is due to the balanced corpus we have created. In the column *diff.*, we show the difference between the 50% baseline and the f-score of the category. As can be seen the f-scores reaches 90%, which is with 40% more than the baseline. According to the $z'$ statistics with confidence level of 0.975, the improvement over the baseline is statistically significant.

| SPANISH | | | | | | |
|---|---|---|---|---|---|---|
| cont/ex | Category | P. | R. | A. | F. | diff. |
| 50/100 | PRESIDENT | 90.38 | 87.67 | 88.83 | 89.00 | |
| | SINGER | 87.94 | 90.00 | 88.33 | 88.96 | +39.00 |
| 50/200 | PRESIDENT | 90.10 | 94.33 | 91.92 | 92.18 | |
| | SINGER | 94.04 | 89.50 | 91.91 | 91.71 | +42.00 |

Table 2: Spanish NE categorization

During the error analysis, we found out that the `PERSON_PRESIDENT` and `PERSON_SINGER` categories are distinguishable and separable because of the well-established semantic similarity relation among the words with which the NE occurs.

A pair of president sentences has lots of strongly related words such as *president:meeting*, *president:governmen*t, which indicates high text cohesion, while the majority of words in a president–singer pair are weakly related, for instance *president:famous*, *president:concert*. But still we found out ambiguous pairs such as *president:company*, where the president relates to a president of a country, while the company refers to a musical enter-

| name | c10 | c25 | c50 | c100 |
|---|---|---|---|---|
| Madonna | **63.63** | **61.61** | 63.16 | **79.45** |
| Julio Iglesias | 58.96 | 56.68 | 66.00 | 79.19 |
| Enrique Iglesias | **77.27** | **80.17** | **84.36** | **90.54** |
| Bill Clinton | 52.72 | 48.81 | **74.74** | 73.91 |
| George Bush | 49.45 | 41.38 | 60.20 | 67.90 |
| Fidel Castro | **61.20** | **62.44** | **77.08** | **82.41** |

Table 3: Spanish NE discrimination

prize. Such information confuses LSA's categorization process and decreases the NE categorization performance.

### 4.3 Spanish name discrimination

In a continuation, we present in Table 3 the f-scores for the Spanish NE discrimination task with the 10, 25, 50 and 100 context windows. The results show that the semantic similarity method we employ is very reliable and suitable not only for the NE categorization, but also for the NE discrimination. A baseline which always returns one and the same person name during the NE discrimination task is 17%. From the table can be seen that all names outperform this baseline. The f-score performance per individual name ranges from 42% to 90%. The results are very good, as the conflated names (three presidents and three singers) can be easily obfuscated, because they share the same domain and occur with the same semantically related words.

The three best discriminated names are Enrique Iglesias, Fidel Castro and Madonna. The name Fidel Castro is easily discriminated due to its characterizing words *Cuba*, *CIA*, *Cuban president*, *revolution*, *tyrant*. All sentences having these words or synonyms related to them are associated to Fidel Castro.

Bill Clinton occurred many times with the words *democracy*, *Boris Yeltsin*, *Halifax*, *Chelsea* (the daughter of Bill Clinton), *White House*, while George Bush appeared with *republican*, *Ronald Reigan*, *Pentagon*, *war in Vietnam*, *Barbara Bush* (the wife of George Bush).

During the data compilation process, the examples for Enrique Iglesias are considered to belong to the Spanish singer. However, in reality some examples of Enrique Iglesias talked about the president of a financial company in Uruguay or political issues. Therefore, this name was confused with Bill Clinton, because they shared semantically related words such as *bank*, *general secretary*, *meeting*, *decision*, *appointment*.

The discrimination process for the singer names is good, though Madonna and Julio Iglesias appeared in the context of *concerts*, *famous*, *artist*, *magazine*, *scene*, *backstage*. The characterizing words for Julio Iglesias are *Chabeli* (the daughter of Julio Iglesias), *Spanish*, *Madrid*, *Iberoamerican*. The name Madonna occurred with words related to a picture of Madonna, a statue in a church of Madonna, the movie Evita.

Looking at the effect of the context window for the NE discrimination task, it can be seen that the best performances of 90% for Enrique Iglesias, 82% for Fidel Castro and 79% for Madonna are achieved with 100 words from the left and from the right of the NE. This shows that the larger context has better discrimination power.

### 4.4 Discussion

After the error analysis, we saw that the performance of our approach depends on the quality of the data source we worked with. Although, we have selected names with low degree of ambiguity, during the data compilation process for which we assumed that they refer 100% to the `SINGER` or `PRESIDENT` categories, during the experiments we found out that one and the same name can refer to three different individuals. This was the case of Madonna and Enrique Iglesias. From one side this impeded the fine-grained categorization and discrimination processes, but opened a new line for research.

In conclusion, the conducted experiments revealed a series of important observations. The first one is that the LSA's term-sentence approach performs better with a higher number of examples, because they provide more semantic information. In addition to the number of examples, the experiments show that the influence of the context window for the name discrimination is significant. The discrimination power is better for larger context windows and this is also related to the expressiveness of the language.

Second, our name categorization and discrimination approach outperforms the baseline with 30%. Finally, LSA is a very appropriate approximation for the resolution of the NE categorization and discrimination tasks. LSA also gives logical explanation about the classification decision of the person names, providing a set of words characterizing the category or simply a list of words describing the individual we want to classify.

## 5 Adaptation to Bulgarian

### 5.1 Motivation

So far, we have discussed and described the development and the performance of our approach with the Spanish language. The obtained results and observations, serve as a base for the context extraction and the experimental setup for the rest of the languages which we want to study. However, to verify the multilingual performance of the approach, we decided to carry out an experiment with a language which is very different from the Romance family.

For this reason, we choose the Bulgarian language, which is the earliest written Slavic language. It dates back from the creation of the old Bulgarian alphabet Glagolista, which was later replaced by the Cyrillic alphabet. The most typical characteristics of the Bulgarian language are the elimination of noun declension, suffixed definite article, lack of a verb infinitive and complicated verb system.

The Bulgarian name discrimination data is extracted from the news corpus Sega2002. This corpus is originally prepared and used in the CLEF competitions. The corpus consists of news articles organized in different XML files depending on the year, month, and day of the publication of the news. We merged all files into a single one, and considered only the text between the text tags. In order to ease the text processing and to avoid encoding problems, we transliterated the Cyrillic characters into Latin ones.

The discrimination data in this experiment consists of the city, country, party, river and mountain categories. We were interested in studying not only the multilingual issue of our approach, but also how scalable it is with other categories. The majority of the categories are locations and only one corresponds to organization. In Table 4, we shows the number of names which we extracted for each one of the categories.

### 5.2 Bulgarian data

The cities include the capital of Bulgaria – Sofia, the second and third biggest Bulgarian cities – Plovdiv and Varna, a city from the southern parts of Bulgaria – Haskovo, the capital of England – London and the capital of Russia – Moskva. The occurrences of these examples are conflated in the ambiguous name `CITY`.

For countries we choose Russia (Rusiya)[3], Germany (Germaniya), France (Franciya), Turkey (Turciya) and England (Angliya). The five names are conflated into `COUNTRY`.

The organizations we worked with are the two leading Bulgarian political parties. BSP (Balgarska Socialisticeska Partija, or Bulgarian Socialist Party) is the left leaning party and the successor to the Bulgarian Communist Party. SDS (Sayuz na demokratichnite sili, or The Union of Democratic Forces) is the right leaning political party. The two organizations are conflated into `PARTY`.

For the `RIVER` category we choose Danube (Dunav) which is the second longest river in Europe and passes by Bulgaria, Maritsa which is the longest river that runs solely in the interior of the Balkans, Struma and Mesta which run in Bulgaria and Greece.

The final category consists of the oldest Bulgarian mountain situated in the southern part of Bulgaria – Rhodope (Rodopi), Rila which is the highest mountain in Bulgaria and on the whole Balkan Peninsula, and Pirin which is the second highest Bulgarian mountain after Rila. The three mountain names are conflated and substituted with the label `MOUNTAIN`.

### 5.3 Bulgarian name discrimination

The experimental settings coincide with those presented in Section 4 and the obtained results are shown in Table 4. The performance of our approach ranges from 32 to 81%. For the five categories, the best performance is achieved for those names that have the majority number of examples.

For instance, for the `CITY` category, the best performance of 79% is reached with Sofia. TAs we have previously mentioned, this is due to the fact that LSA has more evidence about the context in which Sofia appears. It is interesting to note that the city

[3] this is the Bulgarian transliteration for Russia

| Category | Instance | Total | P | R | F |
|---|---|---|---|---|---|
| **City** | Plovdiv | 1822 | 44.42 | 83.87 | 58.08 |
| | **Sofiya** | **5633** | 71.39 | 89.79 | **79.54** |
| | Varna | 1042 | 32.02 | 82.64 | 46.17 |
| | Haskovo | 140 | 21.09 | 69.29 | 32.33 |
| | London | 751 | 31.32 | 84.82 | 45.74 |
| | Moskva | 1087 | 39.47 | 88.22 | 54.53 |
| **Country** | **Rusiya** | **2043** | 55.83 | 86.19 | **67.77** |
| | Germaniya | 1588 | 40.72 | 77.96 | 53.50 |
| | Francia | 1352 | 37.27 | 77.81 | 50.39 |
| | Turciya | 1162 | 43.23 | 84.08 | 57.10 |
| | Angliya | 655 | 29.67 | 72.67 | 42.14 |
| **Party** | BSP | 2323 | 42.54 | 99.35 | 59.57 |
| | **SDS** | **3916** | 64.86 | 98.85 | **78.32** |
| **River** | **Dunav** | **403** | 85.39 | 76.92 | **80.94** |
| | Marica | 203 | 77.88 | 83.25 | 80.47 |
| | Mesta | 81 | 63.64 | 95.06 | 76.24 |
| | Struma | 37 | 56.67 | 91.89 | 70.10 |
| **Mountain** | Rila | 101 | 70.22 | 91.09 | 79.31 |
| | Pirin | 294 | 75.11 | 57.48 | 65.12 |
| | **Rodopi** | **135** | 71.04 | 96.29 | **81.76** |

Table 4: Bulgarian NE discrimination

Varna forms part of weak named entities such as the University of Varna, the Major house of Varna. Although, this strong entity is embedded into the weak ones, practically Varna changes its semantic category from a city into university, major house. This creates additional ambiguity in our already conflated and ambiguous names. In order to improve the performance, we need a better data generation process where the mixture of weak and strong entities will be avoided.

The same effect of best classification for majority sense is observed with the `COUNTRY` category. The best performance of 67% is obtained for Russia. The other country which is distinguished significantly well is Turkey. The 57% performance is from 5 to 10% higher compared to the performances of Germany, England and France. This is due to the context in which the names occur. Turkey is related to trading with Bulgaria and emigration, meanwhile the other countries appear in the context of the European Union, the visit of the Bulgarian president in these countries.

During the error analysis, we noticed that in the context of the political parties, SDS appeared many times in with the names of the political leader or the representatives of the BSP party and vice versa. This impeded LSA's classification, because of the similar context.

Among all categories, `RIVER` and `MOUNTAIN` obtained the best performances. The rivers Dunav and Maritsa reached 80%, while the mountains Rodopi achieved 81.76% f-score. Looking at the discrimination results for the other names in these categories, it can be seen that their performances are much higher compared to the names of the `CITY`, `COUNTY` and `PARTY` categories. This experiment shows that the discrimination power is related to the type of the NE category we want to resolve.

## 6 Conclusions

In this paper, we have presented a language independent approach for person name categorization and discrimination. This approach is based on the sentence semantic similarity information derived from LSA. The approach is evaluated with different NE examples for the Spanish and Bulgarian languages. We have observed the discrimination performance of LSA not only with the `SINGER` and `PRESIDENT` companies, but also with the `CITY`, `COUNTRY`, `MOUNTAIN`, `RIVER` and `PARTY`. This is the first approach which focuses on the resolution of these categories for the Bulgarian language.

The obtained results both for Spanish and Bulgarian are very promising. The baselines are outperformed with 25%. The person fine-grained categorization reaches 90% while the name discrimination varies from 42% to 90%. This variability is related to the degree of the name ambiguity among the conflated names and similar behaviour is observed in the co-occurence approach of (Pedersen et al., 2005).

During the experimental evaluation, we found out that the 100% name purity (e.g. that one name belongs only to one and the same semantic category) which we accept during the data creation in reality contains 9% noise. These observations are confirmed in the additional experimental study we have conducted with the Bulgarian language. According to the obtained results, our text semantic similarity approach performs very well and practically there is no restrain to be adapted to other languages, data sets or even new categories.

## 7 Future Work

In the future, we want to relate the name discrimination and categorization processes, by first encountering the different underlying meanings of a name and then grouping together the sentences that belong to the same semantic category. This process will increase the performance of the NE fine-grained categorization, and will reduce the errors we encountered during the classification of the singers Enrique Iglesias and Madonna. In addition to this experiment, we want to cluster web pages on the basis of name ambiguity. For instance, we want to process the result for the Google's query George Miller, and form three separate clusters obtained on the basis of a fine-grained and name discrimination. Thus we can form the clusters for George Miller the congressman, the movie director and the father of WordNet. This study will include also techniques for automatic cluster stopping.

Moreover, LSA's ability of language independence can be exploited to resolve cross-language NE categorization and discrimination from which we can extract cross-language pairs of semantically related words characterizing a person e.g. George Bush is seen with White House in English, la Casa Blanca in Spanish, a Casa Branka in Portuguese and Beliat Dom in Bulgarian.

With LSA, we can also observe the time consistency property of a person which changes its semantic category across time. For instance, a student turns into a PhD student, teaching assistant and then university professor, or as in the case of Arnold Schwarzenegger from actor to governor.

## Acknowledgements

We would like to thank the three anonymous reviewers for their useful comments and suggestions. This work was partially funded by the European Union under the project QALLME number FP6 IST-033860 and by the Spanish Ministry of Science and Technology under the project TEX-MESS number TIN2006-15265-C06-01.

## References

A. Bagga and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the Thirty-Sixth Annual Meeting of the ACL and Seventeenth International Conference on Computational Linguistics*, pages 79–85.

G. Cleuziou, L. Martin, and C. Vrain. 2004. Poboc: An overlapping clustering algorithm, application to rule-based classification and textual data. In *ECAI*, pages 440–444.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. In *Journal of the American Society for Information Science*, volume 41, pages 391–407.

S. Dumais. 1995. Using lsi for information filtering: Trec-3 experiments. In *The Third Text Retrieval Conference (TREC-3)*, pages 219–230.

M. Fleischman and E. Hovy. 2002. Fine grained classification of named entities. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7.

Z. Kozareva, O. Ferrández, A. Montoyo, R. Muñoz, A. Suárez, and J. Gómez. 2007. Combining data-driven systems for improving named entity recognition. *Data and Knowledge Engineering*, 61(3):449–466, June.

G. Mann. 2002. Fine-grained proper noun ontologies for question answering. In *COLING-02 on SEMANET*, pages 1–7.

G. Miller and W. Charles. 1991. Contextual correlates of semantic similarity. In *Language and Cognitive Processes*, pages 1–28.

M. Pasca. 2004. Acquisition of categorized named entities for web search. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 137–145.

T. Pedersen, A. Purandare, and A. Kulkarni. 2005. Name discrimination by clustering similar contexts. In *CICLing*, pages 226–237.

P. Shah, D. Schneider, C. Matuszek, R.C. Kahlert, B. Aldag, D. Baxter, J. Cabral, M. Witbrock, and J. Curtis. 2006. Automated population of cyc: Extracting information about named-entities from the web. In *Proceedings of the Nineteenth International FLAIRS Conference*, pages 153–158.

H. Shütze. 1998. Automatic word sense discrimination. In *Journal of computational linguistics*, volume 24.

H. Tanev and B. Magnini. 2006. Weakly supervised approaches for ontology population. In *Proceeding of 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–24.