recibido 18-05-2007; aceptado 22-06-2007

The Influence of Context during the Categorization and Discrimination of Spanish and Portuguese Person Names

Zornitsa Kozareva, Sonia Vázquez and Andrés Montoyo
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
zkozareva,svazquez,montoyo@dlsi.ua.es

Resumen: Este artículo presenta un nuevo método para la categorización y la discriminación de nombres propios utilizando como fuente de información la similitud semántica. Para establecer las relaciones semánticas entre las palabras que forman el contexto donde aparece la entidad que queremos categorizar o discriminar, nuestro método utiliza la semántica latente. Se han realizado diferentes experimentos donde se ha estudiado la influencia del contexto y la robustez de nuestra aproximación sobre distintos números de ejemplos. La evaluación se ha realizado sobre textos en español y portugués. Los resultados obteniendos son 90 % para español y 82 % para portugués en categorización y un 80 % para español y un 65 % para portugués en discriminación.
Palabras clave: discriminación de nombres, categorización de nombres, información semántica

Abstract: This paper presents a method for fine-grained categorization and discrimination of person names on the basis of the semantic similarity information. We employ latent semantic analysis which establishes the semantic relations between the words of the context in which the named entities appear. We carry out several experimental studies in which we observe the influence of the context and the robustness of our approach with different number of examples. Our approach is evaluated with Spanish and Portuguese. The experimental results are encouraging, reaching 90 % for the Spanish and 82 % for the Portuguese person name categorization, and 80 % for the Spanish and 65 % for the Portuguese NE discrimination of six conflated names.
Keywords: name discrimination, name categorization, semantic information

1. Introduction and Related Work

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 of drawbacks related to this fine-grained NE issue. First, the systems need hand annotated data which is not available and its creation is time-consuming and requires supervision by experts. Second, for languages other than English there is a significant lack of freely available or developed resources.

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

Together with the specialized NE catego-

ISSN: 1135-5948

rization, we face the problem of name ambiguity which is related to queries for different people, locations or companies that share the same name. This problem is known as name discrimination (Ted Pedersen y Kulkarni, 2005). For instance, Cambridge is a city in United Kingdom, but also in the United States of America. ACL refers to "The Association of Computational Linguistics", "The Association of Christian Librarians", "Automotive Components Limited" among others.

Previously, (Ted Pedersen y Kulkarni, 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 y 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 y 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 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 y Magnini, 2006) classified geographic location and person names into several subclasses. They use syntactic information and observed how often a syntactic pattern cooccurs 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 patters 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.

However, none of these approaches studied the text cohesion and semantic similarity between snippets with named entities. Therefore, we employ Latent Semantic Analysis (LSA) which allows us to establish the semantic relations among the words that surround the named entity. Our motivation is based on the words sense discrimination hypothesis of (Miller y Charles, 1991) according to which words with similar meaning are used in similar context. For instance, names that belong to the category *sport* will be more likely to appear with words such as *championship*, *ball*, *team*, meanwhile names of *university students* or *professors* will be more likely to appear with words such as *book*, *library*, *homework*.

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 (Scott Deerwester y Harshman, 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 analyzes, 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, the NLP applications extract a list of semantically related word pairs or rank documents related to the same topic.

LSA represents explicitly terms and documents in a rich, high 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 dimensionality of the term-document space, 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 sentences in which the NE have to be categorized or discriminated, while the columns correspond to sentences with different named entities. The cells of the matrix show the number of times a given word occurs in a given sentence. 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 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 name categories co-occur and relate semantically to different words.

3. Named Entity Data Set

In order to evaluate our method, we have used two languages: Spanish and Portuguese. We collected large news corpora from the same time period for both languages and identified a predefined set of named entities on the basis of machine-learning based named entity recognizer (Zornitsa Kozareva y Gómez, 2007). The Spanish corpus we worked with is EFE94-95, containing 127079110 tokens. The Portuguese corpora are Folha94-95 and Publico94-95, containing 90809250 tokens. These corpora were previously used in the CLEF competitions[1].

For the NE categorization and discrimination experiments, we used six different low ambiguous named entities, which we assume a-priory to belong to one of the two fine-grained NE categories PERSON_SINGER and PERSON_PRESIDENT. The president names, both for Spanish and Portuguese are Bill Clinton, George Bush and Fidel Castro. The singers for Spanish are Madonna, Julio Iglesias and Enrique Iglesias, while for Portuguese we have Michael Jackson, Madonna and Pedro Abrunhosa. Although we wanted to use the same singer names for both languages, it was impossible due to the scatteredness in the example distribution.

Table 1 shows the original distribution of the extracted examples with different context windows that surround the named entity. The context windows we worked with are 10, 25, 50 and 100. They indicate the number of words[2] from the left and from the right of the identified named entity. Note, that the NE data is obtained only from the content between the text tags in the *xml* documents. During the creation of the context windows, we used 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, the domain and the topic in which the NE is seen can change. Therefore, NE examples for which the number of words from the left or from the right did not correspond to the number of context words were directly discarded.

To avoid imbalance in the experimental data during the evaluation, we decided to create two samples, one with 100 and another with 200 examples per named entity. Thus, every name will have the same frequency of occurrence and there will be no dominance during the identification of a given name.

For the NE categorization data, each occurrence of the president and singer names is replaced with the obfuscated form President_Singer, while for the NE discrimination task, the names where replaces with M_EI_JI_BC_GB_FC. The first label indicates that a given sentence can belong to the president or to the singer category, while the sec-

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

[2] 10, 25, 50 and 100 respectively

name	lang	c10	c25	c50	c100
M	ES	280	266	245	206
	PT	1008	975	893	758
JI	ES	426	405	367	295
EI	ES	407	392	360	305
MJ	PT	592	568	506	418
PA	PT	364	347	320	275
BC	ES	6928	5970	5271	5185
	PT	3055	2951	2786	2576
GB	ES	730	649	641	521
	PT	307	300	283	242
FC	ES	2865	2765	2779	2357
	PT	3050	2951	2777	2460

Table 1: NE distribution in the Spanish and Portuguese corpora

ond 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 named category until the evaluation stage.

4. *Experimental Evaluation*

To carry out the various experimental evaluations, first we construct the conceptual matrix and establish the semantic similarity relations among the sentences in the data set. For each sentence, LSA produces a list of the similarity between all sentences and the target one e.g. the sentence to be classified. The list is ordered in descending order, where high probability values indicate strong similarity and cohesion between the text of the two sentences and vice versa. Therefore, we consider only the top twenty high-scoring sentences, since their NEs will be very likely to belong to the same fine-grained category or person.

In order to evaluate the performance of our approach, we use the standard precision, recall, f-score and accuracy measures which can be derived from Table 2.

number of	Correct `PRES.`	Correct `SING.`
assigned `PRES.`	a	b
assigned `SING.`	c	d

Table 2: Contingency table

$$Accuracy = \frac{a+d}{a+b+c+d} \qquad (1)$$

$$F_{\beta=1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (2)$$

For the assignment of the president and singer categories, we took LSA's list and grouped together in a cluster all sentences from the 20 most similar ones. In contrast, for the NE discrimination task, we did not use the whole list of returned sentences, since we were interested in concrete NE with identical features and characteristics. For this reason, we decided that the most relevant information is contained in the first sentences at the top of LSA's list and rejected the rest of the candidates. The information about the named category or class was not revealed and used until evaluation.

Our experiments are ordered according to the conducted observations. The first one concerns the effect of the context for the NE categorization. This information is very important and beneficial, when annotated corpus has to be created. In this way we can save time and labor for human annotators, or can ease the supervision process after active learning or bootstrapping (Kozareva, 2006). Then, we observe the NE fine-grained classification and discrimination.

4.1. Influence of context

Figures 1 and 2 present the performance of our approach with different context windows. The evaluation is carried out with 100 and 200 examples per NE. For both samples and both languages (Spanish and Portuguese), the context windows perform almost the same.

This shows that on average with 2-3 sentences the context in which the name resides can be captured together with the particular words that characterize and co-occurring with the name.

4.2. NE categorization

In Table 3, we show the results for the Spanish and Portuguese NE fine-grained categorization. The detailed results are only for the 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

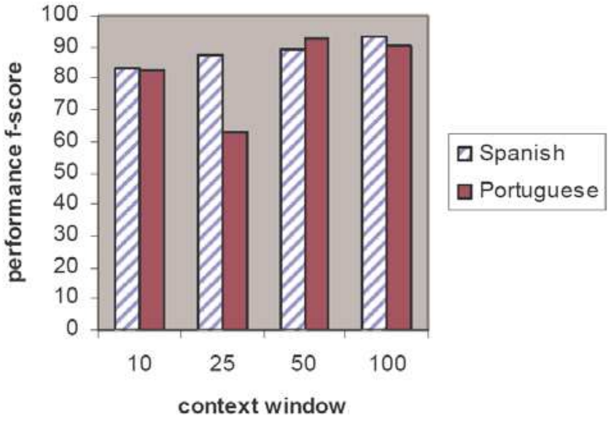

Figure 1: Influence of context for Portuguese and Spanish with 100 examples

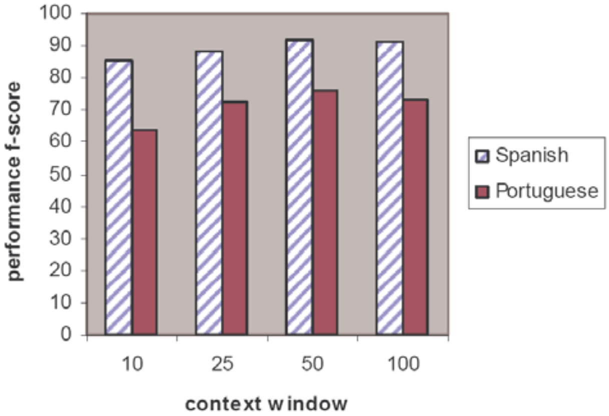

Figure 2: Influence of context for Portuguese and Spanish with 200 examples

performance is due to the balanced corpus we have created. The f-scores for the fine-grained NE categorization in Spanish reach around 90 %, while for Portuguese the f-scored varies around 92 % for the 100 examples, and 76 % for the 200 examples.

SPANISH					
cont/ex	Cat.	P.	R.	A.	F.
50/100	PRES.	90.38	87.67	88.83	89.00
	SING.	87.94	90.00	88.33	88.96
50/200	PRES.	90.10	94.33	91.92	92.18
	SING.	94.04	89.50	91.91	91.71
PORTUGUESE					
cont/ex	Cat.	P.	R.	A.	F.
50/100	PRES.	93.56	92.00	92.50	92.53
	SING.	92.07	56.50	77.17	71.29
50/200	PRES.	96.58	56.50	77.17	71.29
	SING.	69.22	97.83	77.16	81.07

Table 3: NE categorization in Spanish and Portuguese

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 co-occurres. A pair of president sentences has lots of strongly related words such as *president:meeting*, *president:government*, 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 there are ambiguous pairs as *president:company*, where the president relates to a president of a country, while the company refers to a musical enterprize. Such information confuses LSA's categorization process.

4.3. NE discrimination

In a continuation, we present in Table 4 the performance of LSA for the NE discrimination task. The results show that this 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 the baseline. The f-score per individual name ranges from 32 % as the lowest to 90 % as the highest performance. The results are very good, as the conflated names (three presidents and three singers) can be easily obfuscated due to the fact that they share the same domain and co-occur with the same semantically related words.

The three best discriminated names for Spanish are Enrique Iglesias, Fidel Castro and Madonna, while for Portuguese we have Fidel Castro, Bill Clinton and Pedro Abrunhosa. For both languages, the name Fidel Castro was 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 co-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).

Some of the examples for Enrique Igle-

name	lang	10	25	50	100
Madonna	SP	**63.63**	**61.61**	63.16	**79.45**
	PT	59.05	47.37	46.15	55.29
Julio_Iglesias	SP	58.96	56.68	66.00	79.19
Enrique Iglesias	SP	**77.27**	**80.17**	**84.36**	**90.54**
Pedro Abrunhosa	PT	51.26	61.97	*69.63*	*80.17*
Michael Jackson	PT	32.15	*62.64*	48.45	62.07
Bill Clinton	SP	52.72	48.81	**74.74**	73.91
	PT	*60.41*	*73.51*	*64.04*	62.38
George Bush	SP	49.45	41.38	60.20	67.90
	PT	*63.83*	34.07	68.16	*66.67*
Fidel Castro	SP	**61.20**	**62.44**	**77.08**	**82.41**
	PT	*60.64*	*79.79*	*71.61*	*68.26*

Table 4: NE discrimination for Spanish and Portuguese

sias which during the data compiling were assumed as the Spanish singer, in reality talk about the president of a financial company in Uruguay or political issues. Therefore, this name was confused with Bill Clinton as they share semantically related words such as *bank*, *general secretary*, *meeting*, *decision*, *appointment*.

The discrimination process was good though Madonna and Julio Iglesias are singers and appear 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 co-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 for Spanish 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. In comparison for the Portuguese data, the highest coverage of 80 % for Fidel Castro, 73 % for Bill Clinton and 62 % for Michael Jackson are reached with the 25 word window. For the Spanish data, the larger context had better discrimination power, while for Portuguese the more local context was better.

The error analysis shows that the performance of our method depends on the quality of the data source we work with. As there is no hand-annotated NE categorization and discrimination corpora, we had to develop our own corpus by choosing low ambiguous and well known named entities. Even though, during our experiments we found out that one and the same name refers to three different individuals. From one side this made it difficult for the categorization and discrimination processes, but opens new line for research.

In conclusion, the conducted experiments revealed a series of important observations. The first one is that the different context windows perform the same. However, for Spanish better classification is obtained with larger contexts, because this is related to the expressiveness of the Spanish language. Second, we can claim that LSA is a very appropriate approximation for the resolution of the NE categorization and discrimination tasks. Apart it gives logical explanation about the classification decision of the person names giving a set of words characterizing the individual persons or their fine-grained categories.

5. *Conclusions and Work in Progress*

In this paper, we present an approach for NE categorization and discrimination, which is based on semantic similarity information derived from LSA. The approach is evaluated with six different low ambiguous person names, and around 3600 different examples for the Spanish and Portuguese languages. The obtained results are very good and outperform with 15 % the already developed approximations. For the president and singer NE categorization, LSA obtains 90 %, while for the NE discrimination, the results vary from 46 % to 90 % depending on the person name. The variability in the name discrimination power is related to the degree of the name ambiguity. 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 from 5 to 9 % noise.

In (Zornitsa Kozareva y Montoyo, 2007a), we have evaluated the performance of the same approach but for the Bulgarian language. This proves that the approach is language independent, because it only needs a set of context with ambiguous names. In this experimental study, we have focused not only on the multilingual issues but also on the discrimination and classification of names from the location and organization categories. The obtained results demonstrate that the best performance is obtained with the context of 50 words and the easiest category is the location one which includes cities, mountains, rivers and countries. In general, the most difficult classification was for the organization names.

In additional experimental study of (Zornitsa Kozareva y Montoyo, 2007b), we have demonstrated that the combination of the name disambiguation and fine-grained categorization processes can improve the quality of the data needed for the evaluation of our approach.

In the future, we want to resolve cross-language NE discrimination and classification. We are interested in extracting pairs of words that describe and represent the concept of a fine-grained category such as president or a singer and in this way identify new candidates for these categories. We will relate this process with an automatic population of an ontology. Finally, we want to relate this approach with our web people search approximation (Zornitsa Kozareva y Montoyo, 2007c) in order to improve the identification of the name ambiguity detection on the web.

Acknowledgements

This research has been funded by QALLME number FP6 IST-033860 and TEXT-MESS number TIN2006-15265-C06-01.

References

Bagga, Amit y Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. En *Proceedings of the Thirty-Sixth Annual Meeting of the ACL and Seventeenth International Conference on Computational Linguistics*, páginas 79–85.

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

Fleischman, Michael y Eduard Hovy. 2002. Fine grained classification of named entities. En *Proceedings of the 19th international conference on Computational linguistics*, páginas 1–7.

Kozareva, Zornitsa. 2006. Bootstrapping spanish named entities with automatically generated gazetteers. En *Proceedings of EACL*, páginas 17–25.

Mann, Gideon. 2002. Fine-grained proper noun ontologies for question answering. En *COLING-02 on SEMANET*, páginas 1–7.

Miller, George y Walter Charles. 1991. Contextual correlates of semantic similarity. En *Language and Cognitive Processes*, páginas 1–28.

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

Scott Deerwester, Susan Dumais, George Furnas Thomas Landauer y Richard Harshman. 1990. Indexing by latent semantic analysis. En *Journal of the American Society for Information Science*, volumen 41, páginas 391–407.

Shütze, H. 1998. Automatic word sense discrimination. En *Journal of computational linguistics*, volumen 24.

Tanev, Hristo y Bernardo Magnini. 2006. Weakly supervised approaches for ontology population. En *Proceeding of 11th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 17–24.

Ted Pedersen, Amruta Purandare y Anagha Kulkarni. 2005. Name discrimination by clustering similar contexts. En *CICLing*, páginas 226–237.

Zornitsa Kozareva, Óscar Ferrández, Andrés Montoyo Rafael Muñoz Armando Suárez y Jaime Gómez. 2007. Combining data-driven systems for improving named entity recognition. *Data Knowl. Eng.*, 61(3):449–466.

Zornitsa Kozareva, Sonia Vazquez y Andres Montoyo. 2007a. A Language Independent Approach for Name Categorization and Discrimination. En *Proceedings of the ACL 2007 Workshop on Balto-Slavonic Natural Language Processing.*

Zornitsa Kozareva, Sonia Vazquez y Andres Montoyo. 2007b. Discovering the Underlying Meanings and Categories of a Name through Domain and Semantic Information. En *Proceedings of Recent Advances in Natural Language Processing.*

Zornitsa Kozareva, Sonia Vazquez y Andres Montoyo. 2007c. UA-ZSA: Web Page Clustering on the basis of Name Disambiguation. . En *Proceedings of the 4th International Workshop on Semantic Evaluations.*