Fine-Grained Classification of Named Entities Exploiting Latent Semantic Kernels

Claudio Giuliano
FBK-irst
I-38100, Trento, Italy
giuliano@fbk.eu

Abstract

We present a kernel-based approach for fine-grained classification of named entities. The only training data for our algorithm is a few manually annotated entities for each class. We defined kernel functions that implicitly map entities, represented by aggregating all contexts in which they occur, into a latent semantic space derived from Wikipedia. Our method achieves a significant improvement over the state of the art for the task of populating an ontology of people, although requiring considerably less training instances than previous approaches.

1 Introduction

Populating an ontology with relevant entities extracted from unstructured textual documents is a crucial step in Semantic Web and knowledge management systems. As the concepts in an ontology are generally arranged in deep class/subclass hierarchies, the problem of populating ontologies is typically solved top-down, firstly identifying and classifying entities in the most general concepts, and then refining the classification process.

Recent advances have made supervised approaches very successful in entity identification and classification. However, to achieve satisfactory performance, supervised systems must be supplied with a sufficiently large amount of training data, usually consisting of hand tagged texts. As domain specific ontologies generally contains hundreds of subcategories, such approaches are not directly applicable for a more fine-grained categorization because the number of documents required to find sufficient positive examples for all subclasses becomes too large, making the manual annotation very expensive.

Consequently, in the literature, supervised approaches are confined to classify entities into broad categories, such as persons, locations, and organizations, while the fine-grained classification has been approached with minimally supervised (e.g., Tanev and Magnini (2006) and Giuliano and Gliozzo (2008)) and unsupervised learning algorithms (e.g., Cimiano and Völker (2005) and Giuliano and Gliozzo (2007)).

Following this trend, we present a minimally supervised approach to fine-grained categorization of named entities previously recognized into coarse-grained categories, e.g., by a named-entity recognizer. The only training data for our algorithm is a few manually annotated entities for each class. For example, *Niels Bohr*, *Albert Einstein*, and *Enrico Fermi* might be used as examples for the class *physicists*. In some cases, training entities can be acquired (semi-) automatically from existing ontologies allowing us to automatically derive training entities for use with our machine learning algorithm. For instance, we may easily obtain tens of training entities for very specific classes, such as *astronomers*, *materials scientists*, *nuclear physicists*, by querying the Yago ontology (Suchanek et al., 2008).

We represent the entities using features extracted from the textual contexts in which they occur. Specifically, we use a search engine to collect such contexts from the Web. Throughout this paper, we will refer to such a representation as multi-context representation, in contrast to the single-context rep-

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 201–209,
Boulder, Colorado, June 2009.

resentation in which an entity is categorized using solely features extracted from the local context surrounding it, usually a window of a few words around the entity occurrence. Single-context features are commonly used in named-entity recognition, however to assign very specific categories the local context might not provide sufficient information. For example, in the sentence "Prof. Enrico Fermi discovered a way to induce artificial radiation in heavy elements by shooting neutrons into their atomic nuclei," single-context features such as, the prefix *Prof.* and the capital letters, provides enough evidence that Enrico Fermi is a person and a professor. However, to discover that he is a physicist we need to analyze a wider context, or alternatively multiple ones. Recently, Ganti et al. (2008) has shown that exploiting multi-context information can greatly improve the fine-grained classification of named entities, when compared to methods using single context only.

In order to effectively represent entities' multi-contexts, we extend the traditional vector space model (VSM), offering a way to integrate external semantic information in the classification process by means of latent semantic kernels (Shawe-Taylor and Cristianini, 2004). As a result, we obtain a generalized similarity function between multi-contexts that incorporates semantic relations between terms, automatically learned from unlabeled data. In particular, we use Wikipedia to build the latent semantic space. The underlying idea is that similar named entities tend to have a similar description in Wikipedia. As Wikipedia provides reliable information and it exceeds all other encyclopedias in coverage, it should be a valuable resource for the task of populating an ontology. To validate this hypothesis, we compare this model with one built from a news corpus.

Our approach achieves a significant improvement over the state of the art for the task of populating the People Ontology (Giuliano and Gliozzo, 2008), although requiring considerably less training instances than previous approaches. The task consists in classifying person names into a multi-level taxonomy composed of 21 categories derived from WordNet, making very fine-grained distinctions (e.g., physicists vs. mathematicians). It provides a more realistic and challenging benchmark than the ones previously available (e.g., Tanev and Magnini (2006) and Fleischman and Hovy (2002)), that consider a smaller number of categories arranged in a one-level taxonomy.

2 Entity Representation

The goal of our research is to determine the fine-grained categories of named entities requiring a minimal amount of human supervision.

Our method is based on the common assumption that named entities co-occurring with the same (domain-specific) terms are highly probable to refer to the same categories. For example, *quantum mechanics*, *atomic physics*, and *Nobel Prize in physics* are all terms that bound Niels Bohr and Enrico Fermi to the concept of physics.

To automatically derive features for the training and testing entities we proceed as follows. We pair each entity i with a multi-context m_i obtained by querying a search engine with the entity "i" and merging the first M snippets $s_{i,j}$ returned ($1 \leqslant j \leqslant M$). A multi-context is therefore a fictitious document obtained by aggregating snippets, i.e., summary texts of the search engine result. Formally, $m_i = \cup_{j=1}^{M} s_{i,j}$, where the operator $\cup$ denotes the concatenation of strings. For example, Figure 1 (a) and (b) show some snippets retrieved for "Enrico Fermi" and "Albert Einstein," while $s_1 \cup s_2 \cup s_3$ and $s_4 \cup s_5 \cup s_6$ represent their multi-contexts, respectively.

The following section describes how entities' multi-contexts are embedded into the feature space in order to train a kernel-based classifier.

3 Kernels for Fine-Grained Classification of Entities

The strategy adopted by kernel methods (Shawe-Taylor and Cristianini, 2004; Schölkopf and Smola, 2002) consists of splitting the learning problem in two parts. They first embed the input data in a suitable feature space, and then use a linear algorithm (e.g., the perceptron) to discover nonlinear pattern in the input space. Typically, the mapping is performed implicitly by a so-called *kernel function*. The kernel function is a similarity measure between the input data that depends exclusively on the specific data type and domain. A typical similarity function is the inner product between feature vectors. Characterizing the similarity of the inputs plays a crucial role in

s_1: [Enrico Fermi]$_{PER}$ discovered that many nuclear transformations could be conducted by using neutrons.
s_2: [Enrico Fermi]$_{PER}$ led the manhattan project's effort to create the first man-made and self-sustaining nuclear chain.
s_3: [Enrico Fermi]$_{PER}$ was most noted for his work on the development of the first nuclear reactor.

(a)

s_4: [Albert Einstein]$_{PER}$ did not directly participate in the invention of the atomic bomb.
s_5: [Albert Einstein]$_{PER}$ is one of the most recognized and well-known scientists of the century.
s_6: [Albert Einstein]$_{PER}$ was born at Ulm, in Württemberg, Germany, on March 14, 1879.

(b)

Figure 1: Examples of snippets retrieved for Enrico Fermi (a) and Albert Einstein (b).

determining the success or failure of the learning algorithm, and it is one of the central questions in the field of machine learning.

Formally, the kernel is a function $k : X \times X \rightarrow \mathbb{R}$ that takes as input two data objects (e.g., vectors, texts, parse trees) and outputs a real number characterizing their similarity, with the property that the function is symmetric and positive semi-definite. That is, for all $x_i, x_j \in X$, it satisfies

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (1)$$

where ϕ is an explicit mapping from X to an (inner product) feature space $\mathcal{F}$.

In the next sections, we define and combine different kernel functions that calculate the pairwise similarly between multi-contexts. They are the only domain specific element of our classification system, while the learning algorithm is a general purpose component. Many classifiers can be used with kernels. The most popular ones are perceptron, support vector machines (SVM), and k-nearest neighbor (KNN).

3.1 Bag-of-Words Kernel

The simplest method to estimate the similarity between two multi-contexts is to compute the inner product of their vector representations in the VSM. Formally, we define a space of dimensionality N in which each dimension is associated with one word from the dictionary, and the multi-context m is represented by a row vector

$$\phi(m) = (f(t_1, m), f(t_2, m), \ldots, f(t_N, m)), \quad (2)$$

where the function $f(t_i, m)$ records whether a particular token t_i is used in m. Using this representation we define *bag-of-words kernel* between multi-contexts as

$$K_{BOW}(m_1, m_2) = \langle \phi(m_1), \phi(m_2) \rangle \quad (3)$$

However, the bag-of-words representation does not deal well with lexical variability. To significantly reduce the training set size, we need to map contexts containing semantically equivalent terms into similar feature vectors. To this aim, in the next section, we introduce the class of semantic kernels and show how to define an effective semantic VSM using (unlabeled) external knowledge.

3.2 Semantic Kernels

It has been shown that semantic information is fundamental for improving the accuracy and reducing the amount of training data in many natural language tasks, including fine-grained classification of named entities (Fleischman and Hovy, 2002), question classification (Li and Roth, 2005), text categorization (Giozzo and Strapparava, 2005), word sense disambiguation (Gliozzo et al., 2005).

In the context of kernel methods, semantic information can be integrated considering linear transformations of the type $\tilde{\phi}(c_j) = \phi(c_j)\mathbf{S}$, where $\mathbf{S}$ is a $N \times k$ matrix (Shawe-Taylor and Cristianini, 2004). The matrix $\mathbf{S}$ can be rewritten as $\mathbf{S} = \mathbf{WP}$, where $\mathbf{W}$ is a diagonal matrix determining the word weights, while $\mathbf{P}$ is the *word proximity matrix* capturing the semantic relations between words. The proximity matrix $\mathbf{P}$ can be defined by setting non-zero entries between those words whose semantic relation is inferred from an external source of domain knowledge. The *semantic kernel* takes the general form

$$\tilde{k}(m_i, m_j) = \phi(m_i)\mathbf{SS}'\phi(m_j)' = \tilde{\phi}(m_i)\tilde{\phi}(m_j)'. \quad (4)$$

It follows directly from the explicit construction that Equation 4 defines a valid kernel.

WordNet and manually constructed lists of semantically related words typically provide a simple way to introduce semantic information into the

kernel. To define a semantic kernel from such resources, we could explicitly construct the proximity matrix $\mathbf{P}$ by setting its entries to reflect the semantic proximity between the words i and j in the specific lexical resource. However, we prefer an approach that exploits unlabeled data to automatically build the proximity matrix, defining a language and domain independent approach.

3.2.1 Latent Semantic Kernel

To define a proximity matrix, we look at co-occurrence information in a (large) corpus. Two words are considered semantically related if they frequently co-occur in the same texts. We use singular valued decomposition (SVD) to automatically derive the proximity matrix $\mathbf{\Pi}$ from a corpus, represented by its term-by-document matrix $\mathbf{D}$, where the $\mathbf{D}_{i,j}$ entry gives the frequency of term t_i in document d_j.[1] SVD decomposes the term-by-document matrix $\mathbf{D}$ into three matrixes $\mathbf{D} = \mathbf{U\Sigma V}'$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices (i.e., $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}$) whose columns are the eigenvectors of $\mathbf{DD}'$ and $\mathbf{D}'\mathbf{D}$ respectively, and $\mathbf{\Sigma}$ is the diagonal matrix containing the singular values of $\mathbf{D}$.

Under this setting, we define the proximity matrix $\mathbf{\Pi}$ as follows:

$$\mathbf{\Pi} = \mathbf{U}_k\mathbf{\Sigma}_k, \tag{5}$$

where $\mathbf{U}_k$ is the matrix containing the first k columns of $\mathbf{U}$ and k is the dimensionality of the latent semantic space and can be fixed in advance. By using a small number of dimensions, we can define a very compact representation of the proximity matrix and, consequently, reduce the memory requirements while preserving most of the information.

The matrix $\mathbf{\Pi}$ is used to define a linear transformation $\pi : \mathbb{R}^N \rightarrow \mathbb{R}^k$, that maps the vector $\phi(m_j)$, represented in the standard VSM, into the vector $\tilde{\phi}(m_j)$ in the latent semantic space. Formally, π is defined as follows

$$\pi(\phi(m_j)) = \phi(m_j)(\mathbf{W\Pi}) = \tilde{\phi}(m_j), \tag{6}$$

where $\phi(m_j)$ is a row vector, $\mathbf{W}$ is a $N \times N$ diagonal matrix determining the word weights such that $\mathbf{W}_{i,i} = log(idf(w_i))$, where $idf(w_i)$ is the *inverse document frequency* of w_i.

[1] SVD has been first applied to perform latent semantic analysis of terms and latent semantic indexing of documents in large corpora by Deerwester et al. (1990).

Finally, the *latent semantic kernel* is explicitly defined as follows

$$K_{LS}(m_i, m_j) = \langle \pi(\phi(m_i)), \pi(\phi(m_j)) \rangle, \tag{7}$$

where ϕ is the mapping defined in Equation 2 and π is the linear transformation defined in Equation 6. Note that we have used a series of successive mappings each of which adds some further improvement to the multi-context representation.

3.3 Composite Kernel

Finally, to combine the two representations of multi-contexts, we define the composite kernel as follows

$$K_{BOW}(m_1, m_2) + K_{LS}(m_1, m_2). \tag{8}$$

It follows directly from the explicit construction of the feature space and from closure properties of kernels that it is a valid kernel.

4 Experiments

In this section, we compare performance of different kernel setups and previous approaches on an ontology population task.

4.1 Benchmark

Experiments were carried out on the People Ontology (Giuliano and Gliozzo, 2008). An ontology extracted from WordNet, containing 1,657 distinct person instances arranged in a multi-level taxonomy having 21 fine-grained categories (Figure 2). To provide a formal distinction between classes and instances, required to assign instances to classes, the authors followed the directives defined by Gangemi et al. (2003) for OntoWordNet, in which the informal WordNet semantics is re-engineered in terms of a description logic.

In order to have a fair comparison, we reproduced the same experimental settings used in Giuliano and Gliozzo (2008). The population task is cast as a categorization problem, trying to assign person instances to the most specific category. For each class, the instances were randomly split into two equally sized subsets. One is used for training and the other for test, and vice versa. The reported results are the average performance over these two subsets. When an instance is assigned to a sub-class it is also implicitly assigned to all its super-classes. For instance,

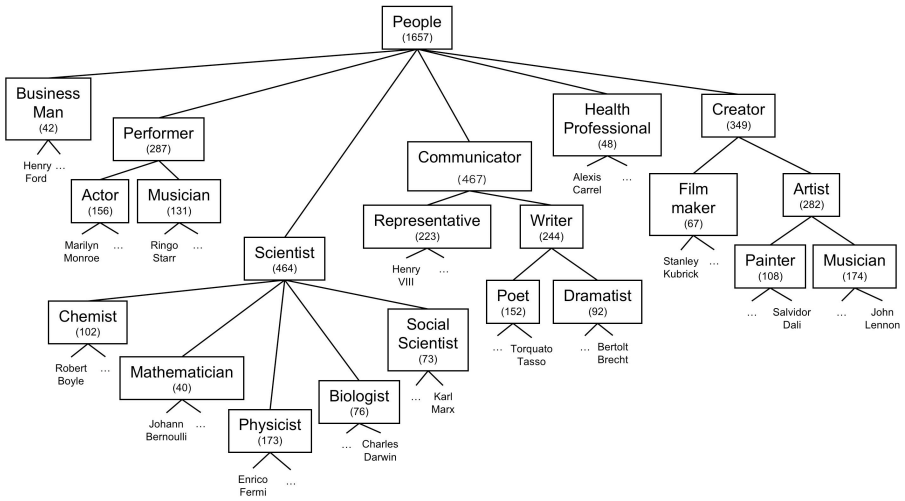

Figure 2: The People Ontology defined by Giuliano and Gliozzo (2008). Numbers in brackets are the total numbers of person instances per category. Concepts with less than 40 instances were removed.

classifying *Salvador Dali* as *painter* we implicitly classify him as *artist* and *creator*. The evaluation is performed as proposed by Melamed and Resnik (2000) for a similar hierarchical categorization task. For instance, classifying *John Lennon* as *painter*, we obtain a false positive for the spurious classification *painter*, a false negative for missing class *musician*, and two true positives for the correct assignment to the super-classes *artist* and *creator*.

4.2 Experimental Settings

We built two proximity matrices Π_W and Π_{NYT}. The former is derived from the 200,000 most visited Wikipedia articles, while the latter from 200,000 articles published by the New York Times between June 1, 1998 and January 01, 2000. After removing terms that occur less than 5 times, the resulting dictionaries contain about 300,000 and 150,000 terms respectively. We used the SVDLIBC package[2] to compute the SVD, truncated to 400 dimensions. To derive the multi-context representation, we collected 100 english snippets for each person instance by querying Google™. To classify each person instance into one of the fine-grained categories, we used a KNN classifier ($K = 1$). No parameter optimization was performed.

[2] `http://tedlab.mit.edu/~dr/svdlibc/`

4.3 Results

Table 1 shows micro- and macro-averaged results for K_{BOW}, K_W, $K_{BOW} + K_W$, K_{NYT}, $K_{BOW} + K_{NYT}$, the IBOP method (Giuliano and Gliozzo, 2008), the random baseline, and most frequent baseline.[3] Where K_W and K_{NYT} are instances of the latent semantic kernel, K_{LS}, using the proximity matrices Π_W and Π_{NYT}, derived from Wikipedia and the New York Times corpus, respectively. Table 2 shows detailed results for each sub- and super-category for $K_{BOW} + K_W$. Table 3 shows the confusion matrix of $K_{BOW} + K_W$, in which the rows are ground truth classes and the columns are predictions. The matrix has been calculated for the finer-grained categories and, then, grouped according to their super-class. To be compared with the IBOP method, all experiments were conducted using only 20 training examples per category. Finally, figure 3 shows the learning curves for $K_{BOW} + K_W$ obtained varying the number of snippets (12, 25, 50, and 100) used to derive the multi-contexts.

4.4 Discussion

On the one hand, the results (Table 2) show that learning the semantic model from Wikipedia gives no significant improvement. Therefore, we reject the hypothesis that encyclopedic knowledge can provide

[3] The most frequent category has been estimated on the training data.

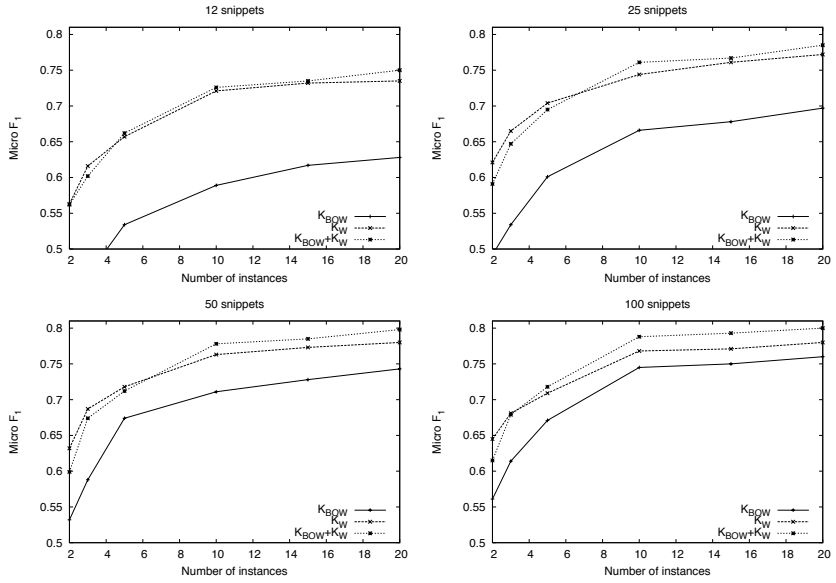

Figure 3: Learning curves for $K_{BOW} + K_W$ obtained varying the number of snippets used to derive the training and test sets. From top-left to bottom right: 12, 25, 50, and 100.

Method	Micro-$\mathbf{F_1}$	Macro-$\mathbf{F_1}$
K_{BOW}	75.6	70.6
K_W	78.1	73.1
$\mathbf{K_{BOW} + K_W}$	**80.0**	**75.4**
K_{NYT}	77.6	72.9
$K_{BOW} + K_{NYT}$	79.7	75.1
IBOP	70.1	62.3
Random	15.4	15.5
Most Frequent	20.7	3.3

Table 1: Comparison among the kernel-based approaches, the IBOP method (Giuliano and Gliozzo, 2008), the random baseline, and most frequent baseline.

more accurate semantic models than general purpose corpora. Moreover, further experiments have shown that even a larger number of Wikipedia articles (600,000) does not help. On the other hand, the latent semantic kernels outperform all the other methods, and their composite ($K_{BOW} + K_W$ and $K_{BOW} + K_{NYT}$) perform the best on every configuration, demonstrating the effectiveness of latent semantic kernels in fine-grained classification of named entities. As in text categorization and word sense disambiguation, they have proven effective tools to overcome the limitation of the VSM by introducing semantic similarity among words.

An important characteristic of the approach is the small number of training examples required per category. This affects both the prediction accuracy and the computation time (this is generally a common property of instance-based algorithms). The learning curves (Figure 3) show that the composite kernel ($K_{BOW} + K_{LS}$) obtained the same performance of the bag-of-word kernel (K_{BOW}) using less than half of the training examples per category. The difference is much more pronounced when using less snippets. The composite kernel $K_{BOW} + K_W$ reaches a plateau around 10 examples, and after 20 examples adding more examples does not significantly improve the classification performance.

As most of entities in the People Ontology are celebrities, all the snippets retrieved by Google™generally refer to them, alleviating the problem of ambiguity of proper names. However, person names are highly ambiguous. In a more realistic scenario, the result of a search engine for a person name is usually a mix of contexts about different entities sharing the same name. In this case, our approach have to be combined with a system that clusters the search engine result, where each cluster is assumed to contain all (and only those) contexts that refer to the same entity. The WePS evaluation campaign on disambiguation of person names (Artiles et al., 2007; Artiles et al., 2009) has shown that the best clustering systems achieve a precision of about 90%

	Scientist					Performer		Creator			Communicator			Business	Health
	Phy	Mat	Che	Bio	Soc	Act	Mus	Fil	Pai	Mus	Poe	Dra	Rep	man	prof
Phy	118	24	10	4	2	0	0	0	0	0	0	0	0	7	2
Mat	2	33	0	0	1	0	0	0	0	0	1	0	0	3	0
Che	13	2	68	9	2	0	0	0	0	0	0	0	0	5	2
Bio	3	0	7	52	0	0	0	0	1	0	0	0	1	6	6
Soc	0	4	1	1	55	0	0	0	0	0	3	1	1	4	2
Act	0	0	0	0	0	98	5	27	0	0	2	14	0	3	0
Mus	0	0	0	0	0	17	67	0	0	32	1	0	1	2	1
Fil	0	0	0	0	0	13	0	45	0	0	1	4	0	2	0
Pai	0	0	0	1	1	2	0	1	100	0	1	0	0	1	0
Mus	0	0	0	0	0	4	29	0	0	139	0	1	0	0	0
Poe	0	2	0	0	0	0	0	0	7	3	98	26	1	2	3
Dra	0	0	0	1	1	9	0	1	0	1	12	61	1	4	1
Rep	0	0	0	0	0	1	1	0	2	0	0	0	197	22	0
Bus	1	0	1	0	1	0	0	1	0	1	0	0	1	36	0
Hea	0	0	0	8	4	0	1	0	0	0	0	1	1	2	31

Table 3: Confusion matrix of $K_{BOW} + K_W$ for the more fine-grained categories grouped according to their top-level concepts of the People Ontology.

Category	**Prec.**	**Recall**	$\mathbf{F_1}$
Scientist	95.1	90.1	92.6
Physicist	86.1	70.7	77.6
Mathematician	50.8	82.5	62.9
Chemist	78.2	67.3	72.3
Biologist	68.4	68.4	68.4
Social scientist	82.1	76.4	79.1
Performer	75.7	69.3	72.3
Actor	68.1	65.8	66.9
Musician	65.0	55.4	59.8
Creator	78.9	82.6	80.7
Film Maker	60.0	69.2	64.3
Artist	83.6	85.4	84.5
Painter	90.9	93.5	92.2
Musician	79.0	80.3	79.7
Communicator	91.9	86.7	89.2
Representative	96.6	88.3	92.3
Writer	86.8	84.2	85.5
Poet	82.4	69.0	75.1
Dramatist	56.5	66.3	61.0
Business man	36.4	85.7	51.1
Health professional	64.6	64.6	64.6
micro	80.9	79.6	80.2
macro	75.1	76.3	75.7

Table 2: Results for each category using $K_{BOW} + K_W$.

and a recall of about 70% and that, in the majority of the cases, the number of contexts per entity is less than 20. This shows that latent semantic kernels are an effective tool for fine-grained classification of person names.

Finally, table 3 shows that misclassification errors are largely distributed among categories belonging to the same super-class (i.e., the blocks on the main diagonal are more densely populated than others). As expected, the algorithm is much more accurate for the top-level concepts (i.e., Scientist, Communicator, etc.), where the category distinctions are clearer, while a further fine-grained classification, in some cases, is even difficult for human annotators.

5 Related Work

Fleischman and Hovy (2002) approach the fine-grained classification of person instances using supervised learning, where the training set is generated semi-automatically, bootstrapping from a small training set. They compare different machine learning algorithms, providing local features as well as global semantic information derived from topic signature and WordNet. Person instances were classified into one of eight categories.

Cimiano and Völker (2005) present an approach for the fine-grained classification of entities relying on the Harris' distributional hypothesis and the vector space model. They assign a particular instance to the most similar concept representing both with lexical-syntactic features extracted from the context of the instance and the lexicalization of the concept, respectively. Experiments were performed using a large ontology with 682 concepts (unfortunately not yet available).

Tanev and Magnini (2006) proposed a weakly-supervised method that requires as training data a list of named entities, without context, for each category under consideration. Given a generic syntactically parsed corpus containing at least each training entity twice, the algorithm learns, for each category,

a feature vector describing the contexts where those entities occur. Then, it compares the new (unknown) entity with the so obtained feature vectors, assigning it to the most similar category. Experiments are performed on a benchmark of 5 sub-classes of location and 5 sub-classes of person.

Giuliano and Gliozzo (2007) propose an unsupervised approach based on lexical entailment, consisting in assigning an entity to the category whose lexicalization can be replaced with its occurrences in a corpus preserving the meaning. Using unsupervised learning, they obtained slightly worst results than Tanev and Magnini (2006) on the same benchmark.

Picca et al. (2007) present an approach for ontology learning from open domain text collections, based on the combination of Super Sense Tagging and Domain Modeling techniques. The system recognizes terms pertinent to the domain and assigns them the correct ontological type.

Giuliano and Gliozzo (2008) present an instance-based learning algorithm for fine-grained named entity classification based on syntactic features (word-order, case-marking, agreement, verb tenses, etc.). Their method can handle much finer distinctions than previous methods, and it is evaluated on a hierarchical taxonomy of 21 ancestors of people that was induced from WordNet. One contribution is to create this richer People Ontology. Another is to make effective use of the Web 1T 5-gram corpus (Brants and Franz, 2006) to represent syntactic information. The main difference between the two approaches lies primarily in the use of syntactic and semantic information. Our experiments show that semantic features do provide richer information than syntactic ones for a more fine-grained classification of named entities. In fact, the accuracy improvement achieved by our approach is more evident for the more specific classes. For example, the improvement in accuracy is about 14% for the class scientist, while it ranges from 25% to 46% for its sub-classes (physicist, mathematician, etc.).

Kozareva et al. (2008) propose an approach for person name categorization based on the domain distribution. They use the information provided by WordNet Domains to generated lists of words relevant for a given domain, by mapping and ranking the words from the WordNet glosses to their WordNet Domains. A named entity is then classified according the similarity between the word-domain lists and the global context in which the entity appears. However, the evaluation was performed only on 6 person names using two categories.

Ganti et al. (2008) present a method that considers an entity's context across multiple documents containing it, and exploiting word n-grams and existing large list of related entities as features. They generated training and test data using Wikipedia articles that contain list of instances. They compare their system with a single-context classifier, showing that their approach based on aggregate context perform better.

Finally, Talukdar et al. (2008) propose a graph-based semi-supervised label propagation algorithm for acquiring open-domain labeled classes and their instances from a combination of unstructured and structured text.

6 Conclusions

We presented an approach to automatic fine-grained categorization of named entities based on kernel methods. Entities are represented by aggregating all contexts in which they occur. We employed latent semantic kernels to extend the bag-of-words representation. The latent semantic models were derived from Wikipedia and a news corpus We evaluated our approach on the People Ontology, a multi-level ontology of people derived from WordNet. Although this benchmark is still far from being "large", it allows drawing more valid conclusions than past ones. We significantly outperformed the previous results on both coarse- and fine-grained classification, although requiring much less training instances. From this preliminary analysis, it appears that semantic information is much more effective that syntactic one for this task, and deriving the semantic model from Wikipedia gives no significant improvement, as well as, using a larger number of Wikipedia articles.

Acknowledgments

Claudio Giuliano is supported by the X-Media project (`http://www.x-media-project.org`), sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978 and the ITCH project (`http://itch.fbk.eu`), sponsored by the Italian Ministry of University and Research and by the Autonomous Province of Trento.

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69, Prague, Czech Republic, June. Association for Computational Linguistics.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2009. Weps 2 evaluation campaign: overview of the web people search clustering task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, Madrid, Spain, April.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1, Linguistic Data Consortium, Philadelphia.

Philipp Cimiano and Johanna Völker. 2005. Towards large-scale, open-domain and ontology-based named entity classification. In *Proceedings of RANLP'05*, pages 66– 166–172, Borovets, Bulgaria.

Scott C. Deerwester, Susan T. Dumais, Thoms K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Michael Fleischman and Eduard Hovy. 2002. Fine grained classification of named entities. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. Axiomatizing WordNet glosses in the OntoWordNet project. In *Proocedings of the Workshop on Human Language Technology for the Semantic Web and Web Services at ISWC 2003*, Sanibel Island, Florida.

Venkatesh Ganti, Arnd C. König, and Rares Vernica. 2008. Entity categorization over large document collections. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 274–282, New York, NY, USA. ACM.

Alfio Giozzo and Carlo Strapparava. 2005. Domain kernels for text categorization. In *Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 56–63, Ann Arbor, Michigan, June.

Claudio Giuliano and Alfio Gliozzo. 2007. Instance based lexical entailment for ontology population. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 248–256.

Claudio Giuliano and Alfio Gliozzo. 2008. Instance-based ontology population exploiting named-entity substitution. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 265–272, Manchester, UK, August.

Alfio Massimiliano Gliozzo, Claudio Giuliano, and Carlo Strapparava. 2005. Domain kernels for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 403–410, Ann Arbor, Michigan, June.

Zornitsa Kozareva, Sonia Vazquez, and Andres Montoyo. 2008. Domain information for fine-grained person name categorization. In *9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008)*, pages 311–321, Haifa, Israel, 17-23 February.

Xin Li and Dan Roth. 2005. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249.

I. Dan Melamed and Philip Resnik. 2000. Tagger evaluation given hierarchical tag sets. *Computers and the Humanities*, pages 79–84.

Davide Picca, Alfio Gliozzo, and Massimiliano Ciaramita. 2007. Semantic domains and supersense tagging for domain-specific ontology learning. In David Evans, Sadaoki Furui, and Chantal Soulé-Dupuy, editors, *Recherche d'Information Assistée par Ordinateur (RIAO)*, Pittsburgh, PA, USA, May.

B. Schölkopf and A. Smola. 2002. *Learning with Kernels*. MIT Press, Cambridge, Massachusetts.

J. Shawe-Taylor and N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A large ontology from wikipedia and wordnet. *Elsevier Journal of Web Semantics*.

Partha Pratim Talukdar, Joseph Reisinger, Marius Pasca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. 2008. Weakly-supervised acquisition of labeled class instances using graph random walks. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, Waikiki, Honolulu, Hawaii, October 25-27.

Hristo Tanev and Bernardo Magnini. 2006. Weakly supervised approaches for ontology population. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.