We are very sorry to have to inform readers of this paper that after publication we have not been able to replicate the evaluation results reported to us by the main author.

Assessing the Challenge of *Fine-Grained* Named Entity Recognition *and* Classification

Asif Ekbal, Eva Sourjikova, Anette Frank and **Simone Paolo Ponzetto**
Department of Computational Linguistics
Heidelberg University, Germany
{ekbal,sourjikova,frank,ponzetto}@cl.uni-heidelberg.de

Abstract

Named Entity Recognition and Classification (NERC) is a well-studied NLP task typically focused on coarse-grained named entity (NE) classes. NERC for more fine-grained semantic NE classes has not been systematically studied. This paper quantifies the difficulty of fine-grained NERC (FG-NERC) when performed at large scale on the people domain. We apply unsupervised acquisition methods to construct a gold standard dataset for FG-NERC. This dataset is used to benchmark methods for classifying NEs at various levels of fine-grainedness using classical NERC techniques and global contextual information inspired from Word Sense Disambiguation approaches. Our results indicate high difficulty of the task and provide a 'strong' baseline for future research.

1 Introduction

Named Entity Recognition and Classification (cf. Nadeau and Sekine (2007)) is a well-established NLP task relevant for nearly all semantic processing and information access applications. NERC has been investigated using supervised (McCallum and Li, 2003), unsupervised (Etzioni et al., 2005) and semi-supervised (Paşca et al., 2006b) learning methods. It has been investigated in multilingual settings (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) and special domains, e.g. biomedicine (Ananiadou et al., 2004).

The classical NERC task is confined to coarse-grained named entity (NE) classes established in the MUC (MUC-7, 1998) or CoNLL (Tjong Kim Sang, 2002) competitions, typically PERS, LOC, ORG, MISC. While most recent work concentrates on feature engineering and robust statistical models for various domains, few researchers addressed the problem of recognizing and categorizing *fine-grained* NE classes (such as *biologist, composer*, or *athlete*) in an open-domain setting.

Fine-grained NERC is expected to be beneficial for a wide spectrum of applications, including Information Retrieval (Mandl and Womser-Hacker, 2005), Information Extraction (Paşca et al., 2006a) or Question-Answering (Pizzato et al., 2006). However, manually compiling wide-coverage gazetteers for fine-grained NE classes is time-consuming and error-prone. Also, without an extrinsic evaluation, it is difficult to define a priori which classes are relevant for a particular domain or task. Finally, prior research in FG-NERC is difficult to evaluate, due to the diversity of NE classes and datasets used.

Accordingly, in the interest of a general approach, we address the challenge of capturing a *broad range* of NE classes *at various levels of conceptual granularity*. By turning FG-NERC into a widely applicable task, applications are free to choose relevant NE categories for specific needs. Also, establishing a gold standard dataset for this task enables comparative benchmarking of methods. However, the envisaged task is far from trivial, given that the set of possible semantic classes for a given NE comprises the full space of NE classes, whereas descriptive nouns may be ambiguous between a fixed set of meanings only.

The paper aims to establish a general framework for FG-NERC by addressing two goals: (i) we automatically build a gold standard dataset of NE instances classified in context with fine-grained semantic class labels; (ii) we develop strong baseline methods, to assess the aptness of standard NLP approaches for this task. The two efforts are strongly interleaved: a standardized dataset is not only essential for (comparative) evaluation, but also a prerequisite for classification approaches based on supervised learning, the most successful techniques for sequential labeling problems.

Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 93–101,
Uppsala, Sweden, 16 July 2010. ©2010 Association for Computational Linguistics

2 Related work

An early approach to FG-NERC is Alfonseca and Manandhar (2002), who identify it as a problem related to Word Sense Disambiguation (WSD). They jointly address concept hierarchy learning and instance classification using topic signatures, yet the experiments are restricted to a small ontology of 9 classes. Similarly, Fleischman and Hovy (2002) extend previous work from Fleischman (2001) on locations and address the acquisition of instances for 8 fine-grained person classes. For supervised training they compile a web corpus which is filtered using high-confident classifications from an initial classifier trained on seeds. Due to the limitations of their method to create a good sample of training data, the performance could not be generalized to held-out data.

Recent work takes the task of FG-NERC one step further by (i) extending the number of classes, (ii) relating them to reference concept hierarchies and (iii) exploring methods for building training and evaluation data, or applying weakly and unsupervised learning based on high-volume data. Tanev and Magnini (2006) selected 10 NE-subclasses of person and location using WordNet as a reference. Datasets were automatically acquired and manually filtered. They compare word and pattern-based supervised and a semi-supervised approach based on syntactic features. Giuliano & Gliozzo (2007, 2008) perform NE classification against the People Ontology, an excerpt of the WordNet hierarchy, comprising 21 people classes populated with at least 40 instances. Using minimally supervised lexical substitution methods, they cast NE classification as an ontology population task – as opposed to recognition and classification in context. In a similar setting, Giuliano (2009) explores semi-supervised classification of the People Ontology classes using latent semantic kernels, comparing models built from Wikipedia and from a news corpus. In a different line of research Paşca (2007) and Paşca and van Durme (2008) make use of query logs to acquire NEs on a large scale. While Paşca (2007) extracts NEs for 10 target classes, Paşca and van Durme (2008) combine web query logs and web documents to acquire both NE-concept pairs and concept attributes using seeds.

But while these more recent approaches all offer substantially novel contributions for many NE acquisition subtasks, none of them addresses the full task of FG-NERC, i.e., recognition and classification of NE tokens *in context*. Compared to ontology population, focusing on types, classification in raw texts needs to consider any token and cannot rely on special contexts offering indicative clues for class membership.

Bunescu and Paşca (2006) also perform disambiguation and classification of NEs in context, yet in a different setup. Disambiguation is performed into one of the known possible classes for a NE, as determined from Wikipedia disambiguation pages. Contexts for training and testing are acquired from Wikipedia pages, as opposed to general text. Disambiguation is performed using vectors of co-occurring terms and a taxonomy-based kernel that integrates word-category correlations. Evaluation is performed on the task of predicting, for a given NE in a Wikipedia page context, the correct class from among its known classes, including one experiment that included 10% of out-of-Wikipedia entities. The category space was confined to *People by occupation*, with 8,202 subclasses. Classification considered 110 broad classes, 540 highly populated classes (w/o out-of-Wikipedia entities), and 2,847 classes including less populated ones. This setup is difficult to compare given the sense granularities employed and the special Wikipedia text genre. Even though classification is performed in context, the task does not evaluate recognition.

To summarize, the field has developed robust methods for acquisition and fine-grained classification of NEs on a large scale. But, the full task of NE recognition and classification in context still remains to be addressed for a wide-coverage, fine-grained semantic class inventory that can serve as a common benchmark for future research.

3 Fine-grained NERC on a large-scale

We present experiments that assess the difficulty of open-domain FG-NERC pursued at a large scale. We concentrate on instances and classes referring to people, since it is a well-studied domain (see Section 2) and structured fine-grained information can be readily applied to a well-defined end-user task such as IR, cf. the Web People Search task (Artiles et al., 2008). Our method is general in that it requires only a (PoS tagged and chunked) corpus and a reference taxonomy to provide a concept hierarchy. Given a mapping between automatically extracted class labels and concepts in a taxonomic resource, it can be further extended to other domains, e.g. locations or the biomedical domain by leveraging open-domain taxonomies such as Yago (Suchanek et al., 2008) or WikiTaxonomy (Ponzetto and Strube, 2007). The contribution of this work is two-fold:

(i) We develop an unsupervised method for acquiring a comprehensive dataset for FG-NERC by applying linguistically motivated patterns to a corpus harvested from the Web (Section 4). Large amounts of NEs are acquired together with their contexts of occurrence and with their fine-grained class labels which are mapped to synsets in WordNet. The controlled sense inventory and the taxonomic structure offered by WordNet enables an evaluation of FG-NERC performance at different levels of concept granularity, as given by the depth at which the concepts are found. As our extraction patterns reflect a wide-spread grammatical construct, the method can be applied to many languages and extended to other domains.

(ii) Given this automatically acquired dataset, we assess the problem of FG-NERC in a systematic series of experiments, exploring the performance of NERC methods on different levels of granularities. For recognition and classification we apply standard sequential labeling techniques – i.e. a Maximum Entropy (MaxEnt) tagger (Section 5.1) – which we adapt to this hierarchical classification problem (Section 5.2). To test the hypothesis of whether a sequential labeler represents a valid choice to perform FG-NERC, we compare the latter to a MaxEnt system trained on a more semantically informed feature set, and a gloss-overlap method inspired by WSD approaches (Section 5.3).

4 Acquisition of a FG-NERC dataset

We present an unsupervised method that simultaneously acquires NEs, their semantic class and contexts of occurrence from large textual resources. In order to develop a clean resource of properly disambiguated NEs, we develop acquisition patterns for a grammatical construction that unambiguously associates proper names with their corresponding semantic class.

Pattern-based extraction of NE-concept pairs. NEs are often introduced by so-called *appositional structures* as in (1), which overtly express which semantic class (here, *painter*) the NE (*Kandinsky*) belongs to. Appositions involving proper names can be captured by extraction patterns as given in (2).

(1) *. . . writings of the abstract painter Kandinsky frequently explored similarities between . . .*

(2) a. [the|The]? [JJ|NN]* [NN] [NP]
the abstract painter Kandinsky

b. [NP] [,]? [a|an|the]* [JJ|NN]* [NN]
W. Kandinsky, a Russian-born painter, ..

Contexts like (2.a) provide a less noisy sequence for extraction, due to the class and instance labels being adjacent – in contrast to (2.b) where any number of modifiers can intervene between the two. Accordingly, we apply in our experiments only a restricted version of (2.a) – with a determiner – to UKWAC, an English web-based corpus (Baroni et al., 2009) that comes in a cleaned, PoS-tagged and lemmatized form. Due to its size (>2 billion tokens) and mixed genres, the corpus is ideally suited for acquiring large quantities of NEs pertaining to a broad variety of open-domain semantic classes.

Filtering heuristics. The apposition patterns are subject to noise, due to PoS-tagging errors, as well as special constructions, e.g. reduced relative clauses. The former can be controlled by frequency filters, the latter can be circumvented by using chunk boundary information[1]. A more challenging problem is to recognize whether an extracted nominal is in fact a valid semantic class for NEs. Besides, class labels can be ambiguous, so there is uncertainty as to which class an extracted entity should be assigned to. We apply two filtering strategies: we set a frequency threshold ft on the number of extracted NE tokens per class, to remove infrequent class label extractions; we then filter invalid semantic classes using information from WordNet: given the WordNet PERSON supersense, i.e. the lexicographer file for nouns denoting people, we check whether the first sense of the class label candidate is found in PERSON.

Mapping to the WordNet person domain. In order to perform a hierarchical classification of people, we need a taxonomy for the domain at hand. We achieve this by mapping the extracted class labels to WordNet synsets. In our setting, we map against all synsets found under *person#n#1*,

[1] We use YamCha (Kudo and Matsumoto, 2000) to perform phrase chunking.

which are direct hypernyms of at least one instance in WordNet ($C_{WN\_pers+Inst}$).[2] Since our goal is to map class labels to synsets (i.e. our future NE classes), we check each class label candidate against all synonyms contained in the synset. At this point we have to deal with two cases: two extracted class label candidates (synonyms such as *doctor, physician*) will map to a single synset, while ambiguous class labels (e.g. *director*) can be mapped to more than one synset. In the latter case, we heuristically choose the synset which dominates the highest number of instances in WordNet.

Mapping evaluation. We evaluated the coverage of our mapping for two sets of class labels extracted for two different frequency thresholds: $ft = 40$ and $ft = 1$. With $ft = 40$, we cover 31.1% of the synsets found under *person#n#1* in WordNet, i.e. the set of classes $C_{WN\_pers+Inst}$; conversely, 45.8% of the extracted class labels can be successfully mapped to $C_{WN\_pers+Inst}$. For threshold $ft = 1$, we are able to map to 87.9% of $C_{WN\_pers+Inst}$, with only 20.1% of extracted classes mapped to $C_{WN\_pers+Inst}$. For the remaining 79.9% of class labels (e.g. *goalkeeper, chancellor, superstar*) that have no instances in WordNet, we manually inspected 20 classes, in 20 contexts each, and established that 76% of them are appropriate NE person classes.

For threshold $ft = 40$, we obtain 153 class labels which are mapped to 146 synsets. Ten class labels are mapped to more than one synset. Using our mapping heuristic based on the majority instance class, we successfully disambiguate all of them. However, since we only map to $C_{WN\_pers+Inst}$, we introduce errors for 5 classes. E.g. ‘manager’ incorrectly gets mapped to *manager#n#2*, since the latter is the only synset containing instances. For these cases we manually corrected the automatic mapping.

A taxonomy for FG-NERC. We create our gold standard taxonomy of semantic classes by starting with the 146 synsets obtained from the mapping, including the 5 classes that were manually corrected. Since we concentrate on the people domain, we additionally remove 5 classes that can refer to other domains as well (e.g. *carrier*, *guide*). Given the remaining 141 synsets, we select the portion of WordNet rooted at *person#n#1* which contains them, together with any intervening synset found along the WordNet hierarchy. Given this WordNet excerpt, the extracted NE tokens are then appended to the respective synsets in the hierarchy. Statistics of the resulting WordNet fragment augmented with instances are given in Table 1. The taxonomy has a maximum depth of 8, and contains 213 synsets, i.e. NE classes (see column 2). 83.5% of the 31,819 extracted instances (type-level) sit in leaf nodes. The classes automatically refer back to the acquired appositional contexts. Table 1 gives statistics about the number of instances (token-level) acquired for classes at different embedding levels. In total we have at our disposal 48,512 instances (token-level) in appositional contexts. The type-token ratio is 1.52.

Level	#C	#C w/inst	#inst	#inst/C	% of inst
1	1	0	0	-	-
2	29	8	2,662	332	5.49
3	57	37	18,229	493	37.58
4	63	46	18,422	401	37.94
5	37	30	6,231	208	12.84
6	18	13	2,366	182	4.88
7	6	5	423	85	0.87
8	2	2	179	90	0.36
all	213	141	48,512	344	100

Table 1: Level-wise statistics of classes and instances across the FG-NERC person taxonomy.

Gold standard validation. To create a gold standard dataset of entities in context labeled with fine-grained classes, we first randomly select 20 classes, as well as an additional 18 which are also found in the People Ontology (Giuliano and Gliozzo, 2008). For each class, we randomly select 40 occurrences of instances in context, i.e. the words co-occurring in a window of 60 tokens before and after the instance. We asked four annotators to label these extractions for correctness, and to provide the correct label for the incorrect cases, if one was available. Only 52 contexts out of 1520 were labeled as incorrect, thus giving us 96.58% accuracy on our automatically extracted data. The manually validated dataset is used to provide a ground-truth for FG-NERC. However, the noun (e.g. *hunter*) denoting the NE class is removed from these contexts for training and testing in all experiments. This is because, due to the extraction method based on POS-patterns denoting appositions, class labels are known *a priori* to occur in the context of an instance and thus identify them with high precision.

[2] We use WordNet version 3.0. With *w#p#i* we denote the i-th sense of a word w with part of speech p. E.g., *person#n#1* is defined as { *person*, *individual* ... }).

5 Methodology for FG-NERC

We develop methods to perform FG-NERC using standard techniques developed for coarse-grained NERC and WSD. These are applied to our dataset from Section 4, in order to measure performance at different levels of semantic class granularity, i.e. corresponding to the depth of the semantic classes found in our WordNet fragment. We start in Section 5.1 to present a Maximum Entropy model to perform coarse-grained NERC and we extend it to perform multiclass classification in a hierarchical taxonomy (Section 5.2). We then present in Section 5.3 an alternative proposal to perform FG-NERC using global context information, as found in state-of-the-art approaches to supervised and unsupervised WSD.

5.1 NERC using a MaxEnt tagger

Our baseline system is modeled following a Maximum Entropy approach (Bender et al., 2003, *inter alia*). The MaxEnt model produces a probability for each class label t (the NE tag) of a classification instance, conditioned on its context of occurrence h. This probability is calculated by:

$$P(t|h) = \frac{1}{Z(h)} \exp\left(\sum_{j=1}^{n} \lambda_j f_j(h,t)\right) \qquad (1)$$

where $f_j(h,t)$ is the j-th feature with associated weight λ_j and $Z(h)$ is a normalization constant to ensure a proper probability distribution.[3] Given a word w_i to be classified as Beginning, Inside or Outside (IOB) of a NE, we extract as features:

1. **Context words**. The words occurring within the context window $w_{i-2}^{i+2} = w_{i-2} \dots w_{i+2}$.
2. **Word prefix and suffix**. Word prefix and suffix character sequences of length up to n.
3. **Infrequent word**. A feature that fires if w_i occurs in the training set less frequently than a given threshold (i.e. below 10 occurrences).
4. **Part-of-Speech (PoS) and chunk information**. The PoS and chunk labels of w_i.
5. **Capitalization**. A binary feature that checks whether w_i starts with a capital letter or not.
6. **Word length**. A binary feature that fires if the length of w_i is smaller than a pre-defined threshold (i.e. less than 5 characters).
7. **Digit and symbol features**. Three features check whether w_i contains digit strings, non-characters (e.g. slashes) or number expressions.
8. **Dynamic feature**. The tag t_{i-1} of the word w_{i-1} preceding w_i in the sequence w_1^n.

5.2 MaxEnt extension for FG-NERC

Extension to hierarchical classification. We apply our baseline NERC system to FG-NERC. Given a word in context, the task consists of recognizing it as a NE, and classifying it into the appropriate semantic class from our person taxonomy. We approach this as a hierarchical classification task by generating a binary classifier[4] with separate training and test sets for each node in the tree.

To perform *level-wise classification* from coarse to fine-grained classes, we need to adjust the class labels and their corresponding training and test instances for each experiment. For classification at the deepest level, each concept contains the instances of the original dataset. For classification at higher levels we leverage the semantics of the WordNet hyponym relations and expand the set of target classes (i.e. synsets) of a given level to contain all instances of hyponym synsets. Given a set I of classification instances for a given target class c, we add all instances labeled with the hyponyms of c to I. All other instances (not in that subtree) are labeled as being Outside (O-) a NE. This approach ensures that, for each node, the dataset contains two classes (NE and O) only, and implicitly 'propagates' the instances up the tree. As a result, non-leaf nodes that did not have any instance in the original dataset become populated. Also, the classification of classes at higher levels is based on larger datasets.

Extension to multiclass classification. Since we train a binary classifier for each node of the tree, we apply two methods to infer multiclass decisions from these binary classifiers, namely *level-wise* and *global* multiclass classification. In both paradigms, we combine the single decisions of the individual classifiers with the winner-takes-all strategy, using weighted voting. The weights are calculated based on the confidence value for the corresponding class, i.e., its conditional probability according to Equation (1). The output label is selected randomly in case of ties.

[3] In our implementation, we use the OpenNLP MaxEnt library (`http://maxent.sourceforge.net`).

[4] The IOB tagging scheme normally assigns three different labels, i.e. Inside (I-), Outside (O-) and Beginning (B-) of a chunk. However, our dataset does not have any instance labeled as B-, since it does not contain any adjacent NEs.

For *level-wise classification*, we combine only classifiers at the same level of embedding. Given n concepts at level l, we have n possible output labels for each word. The output label for a classification instance is determined by the highest weighted vote among all binary classifiers at level l. For *global classification* we combine all binary classifiers of the entire tree using weighted voting to determine the winning class label. The weights are calculated based on the product of confidence value and depth of the corresponding class in the tree.

5.3 FG-NERC using global contexts

FG-NERC is a more demanding task than 'classical' NERC, due to the larger amount of classes, the paucity of examples for each class, and the increasingly subtle semantic differences between these classes. For such a task contextual information is expected to be very informative – e.g. if an entity co-occurs in context with *'Nobel prize'*, this provides evidence that it is likely to be a *scientist* or *scholar*. However, the context window used by our baseline MaxEnt tagger is very local, including at most the two preceding and succeeding words. Hence, the classifier is not able to capture informative contextual clues in a larger context.

Previous work has related FG-NERC to WSD approaches (Alfonseca and Manandhar, 2002). Accordingly, we investigate two context-sensitive approaches inspired from WSD proposals, which consider a more *global context* for classification. We first define a new feature set to induce a new MaxEnt model (MaxEnt-B) which only uses lexical features from a larger context window, as used in standard supervised WSD (Lee and Ng, 2002):

1. **PoS context**. The part-of-speech occurring within the context window $pos_{i-3}^{i+3} = pos_{i-3} \dots pos_{i+3}$.
2. **Local collocation**. Local collocations C_{nm} surrounding w_i. We use $C_{-2,-1}$ and $C_{1,2}$.
3. **Content words in surrounding context**. We consider all unigrams in contexts $w_{i-3}^{i+3} = w_{i-3} \dots w_{i+3}$ of w_i (crossing sentence boundaries) for the entire training data. We convert tokens to lower case, remove stopwords, numbers and punctuation symbols. We define a feature vector of length 10 using the 10 most frequent content words. Given a classification instance, the feature corresponding to token t is set to 1 iff the context w_{i-3}^{i+3} of w_i contains t.

In addition, we use a Lesk-like method (Lesk, 1986) which labels instances in context with the WordNet synset whose gloss has the maximum overlap with the glosses of the senses of its words in context. Given the small context provided by the WordNet glosses, we follow Banerjee and Pedersen (2003) and expand these to also include the words from the glosses of the hypernym and hyponym synsets.

6 Experiments

6.1 Benchmarking on coarse-grained NERC

We benchmark the performance of our baseline MaxEnt classifier using the feature set from Section 5.1 (MaxEnt-A henceforth) on the CoNLL-2003 shared task dataset (Tjong Kim Sang and De Meulder, 2003), the *de-facto* standard for evaluating coarse-grained NERC systems.

In MaxEnt modeling, feature selection is a crucial problem and key to improving classification performance. MaxEnt, however, does not provide methods for automatic feature selection. We therefore experimented with various combinations of features standardly used for NERC (1-8 of Section 5.1). Model parameters are computed with 200 iterations without feature frequency cutoff. The best configuration is found by optimizing the F_1 measure on the development data with various feature representations. The chosen features are: 1, 2 (with $n = 3$), 4, 5, 6, 7 and 8. Evaluation on the test set is performed blindly, using this feature set. The results are presented in Table 2.

The MaxEnt labeler achieves performance comparable with the CoNLL-2003 task participants, ranking 12^{th} among the 16 systems participating in the task, with a 2 point margin off the F_1 of the most similar system of Bender et al. (2003) and 7 points below the best-performing system (Florian et al., 2003). The former used a relatively complex set of features and different gazetteers extracted from unannotated data. The latter combined four diverse classifiers, namely a robust linear classifier, maximum entropy, transformation-based learning and a hidden Markov model. They used different feature sets, unannotated data and an additional NE tagger. In comparison, our NERC system is simpler and based on a small set of features that can be easily obtained for many languages. Besides, it does not make use of any external resources and still shows state-of-the-art performance on the overall data.

	Recall	Precision	$F_{\beta=1}$
PER	83.02%	81.40%	82.21%
LOC	88.47%	88.19%	88.23%
ORG	77.20%	68.03%	72.23%
MISC	81.20%	83.92%	82.54%
Overall	83.11%	80.47%	81.77%

Table 2: Results on the CoNLL-2003 test data.

Set	# tokens	# NEs
Training	2,431,041	38,810
Development	478,871	9,702
Test	181,490	1,520

Table 3: Statistics for training, dev and test sets.

6.2 Evaluating FG-NERC

Experimental setting. For the task of FG-NERC, we compare the performance of MaxEnt-A with the MaxEnt-B model from Section 5.3 and the Lesk method. The data is partitioned into training and development sets by randomly selecting 80%-20% of the contexts in which the NEs occur. We use the held-out, manually validated gold standard from Section 4 for blind evaluation. Statistics for the dataset are reported in Table 3.

We build a MaxEnt model for each FG-NE class, using the features that performed best on the CoNLL task, *except* the digit and dynamic NE features (MaxEnt-A), and context features 1-3 of Section 5.3 (MaxEnt-B). Model parameters are computed in the same way as for coarse-grained NERC. Table 3 shows that our training set is highly unbalanced. The ratio between positive (NEs) and negative examples (i.e. O classification instances) at the topmost level is 63:1. Also, with increasing levels of fine-grainedness, the number of negative (-O) NE classes is increasing for each binary classifier. We observed on the development set that this skewed distribution heavily biases the classifiers towards the negative category, and accordingly investigated sampling techniques to make the ratio of positive and negative examples more balanced. We experiment with a sampling strategy that over-samples the positive examples and under-samples the negative ones. We define various ratios of over-sampling depending upon the number of positive examples in the original training set. Table 4 lists the factors (f) of over-sampling applied to the original positive samples (P), with minimum and maximum sizes of the obtained oversampled sets P' for different ranges of original sizes of P.[5] Oversampling is done without replacement. The number of negative instances is always downsampled on the basis of P' to yield a 1:5 ratio of positive and negative samples, a ratio we estimated from the CoNLL-2003 data.

factor f	size of P	min P'	max P'
$20 \times P$	1 – 2K	20	40K
$15 \times P$	2K – 5K	30K	75K
$10 \times P$	5K – 10K	50K	100K
$5 \times P$	10K – 20K	50K	100K
$2 \times P$	20K – 50K	40K	100K
P	50K – . . .	50K	>50K

Table 4: Oversampling of positive samples.

Level	MaxEnt-A			MaxEnt-B		
	R	P	F_1	R	P	F_1
1	98.7	85.0	91.4	95.1	83.0	88.6
2	96.0	65.5	77.9	48.1	46.3	47.2
3	95.3	54.3	69.3	43.3	41.1	42.2
4	86.8	52.8	65.6	41.1	37.2	39.1
5	90.4	45.9	60.9	49.2	21.5	29.9
6	91.6	36.9	52.6	51.7	13.2	21.1
7	89.5	31.8	46.9	42.2	10.2	16.4
8	100.0	19.9	66.7	87.1	8.1	14.7
global	85.1	43.2	57.3	61.9	26.6	37.2
hierarchical	87.7	44.8	59.4	64.5	29.5	40.5

Table 6: Level-wise NE recognition & classification evaluation (in %).

Level-wise evaluation results on the *FG-NE classification-only (NEC) task* for the MaxEnt classifiers and Lesk are given in Table 5. Table 6 reports results for the evaluation of the MaxEnt model performing *both classification and recognition*. As for coarse-grained NERC, we evaluate using the standard metrics of recall (R), precision (P) and balanced F-measure (F_1). As baseline, we use a majority class assignment – i.e. at each level, we label all instances with the most frequent class label. For *global FG-NE classification*, reported in Table 5, the original fine-grained classes are considered, across the entire class hierarchy. Global evaluation is performed by counting exact label predictions on the entire hierarchy (*global*) and using the evaluation metric of Melamed and Resnik (2000, *hierarchical*). As baseline we assume the most frequent class label in the training set.

Discussion. All methods perform reasonably well, indicating the feasibility of the task. For the MaxEnt models, Table 5 shows a general high recall and decreasing precision as we move down the hierarchy. Degradation in the overall F_1 score is

[5] Sampling ratios are determined on the development set.

Level	Baseline			MaxEnt-A			MaxEnt-B			Lesk		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	28.4	25.9	27.1	85.8	88.6	87.0	79.5	84.9	82.2	16.4	19.7	17.9
3	27.9	23.1	25.2	83.9	88.1	85.9	75.5	79.8	77.5	16.2	16.2	16.2
4	18.8	20.4	19.5	74.6	85.0	79.5	65.4	71.3	68.2	11.3	11.3	11.3
5	25.8	19.0	21.9	78.8	83.4	80.9	78.6	74.1	76.3	13.5	14	13.8
6	24.7	7.8	11.9	88.5	73.6	80.4	78.7	74.1	75.7	33.2	37.5	35.2
7	19.1	5.34	8.3	79.2	76.5	77.8	78.1	72.7	75.3	49.4	49.4	49.4
8	34.2	2.9	5.5	82.8	73.8	78.1	81.1	71.1	75.8	0.1	0.1	0.1
global	34.6	18.5	24.1	81.1	84.2	82.6	78.0	74.2	76.6	36.5	38.6	37.5
hierarchical	33.0	21.2	25.8	83.5	86.2	84.8	78.2	77.8	78.1	36.6	38.7	37.6

Table 5: Level-wise evaluation of fine-grained NE classification techniques (in %).

given by the increasingly limited amount of class instances found towards the low regions of the tree (down to an average of 85 and 90 instances per class for levels 7 and 8, respectively) (cf. Table 1). The 'classical' feature set (MaxEnt-A) yields better performance compared to the semantic feature set (MaxEnt-B). However MaxEnt-B still achieves a respectable performance, given that it contains a few semantic features only.

The MaxEnt classifiers achieve a far better performance than Lesk. This is in-line with previous findings in WSD, namely unsupervised fine-grained disambiguation methods rarely performing above the baseline, and suggests that Lesk can be merely used as a 'strong' baseline. Error analysis showed that it performs poorly due to the limited context provided by the WordNet glosses, and the limited impact of gloss expansions deriving from the low connectivity between synsets.

Comparison of Tables 5 and 6 shows that performance decreases considerably for a classifier that not only assigns fine-grained classes, but also detects which tokens actually are NEs. As for the classification-only task, the performance decreases as one moves to lower levels. This indicates that the complexity of the task is proportional to the fine-grainedness of the class inventory. MaxEnt-B, lacking 'classical' NER features, shows dramatic losses, compared to MaxEnt-A.

Comparison to other work. We compared the performance of our system based on global classification (one vs. rest) against the figures reported for individual categories in Giuliano (2009). The MaxEnt-A system compares favorably, although it considers (i) more classes at each level – i.e. 213 vs. 21 – and (ii) classifies NEs at finer-grained levels – i.e. 8 vs. 4 maximum depth in the respective WordNet fragments. We achieve overall micro average R, P and F_1 values of 87.5%, 85.7% and 86.6%, respectively, compared to Giuliano's 79.6%, 80.9% and 80.2%. Due to the different setups and data used, these figures do not offer a basis for true comparison. However, the figures suggest that our system achieves respectable performance on a more complex classification problem.

7 Conclusions

We presented a method to perform FG-NERC on a large scale. Our contribution lies in the definition of a benchmarking setup for this task in terms of gold standard datasets and strong baseline methods provided by a MaxEnt classifier. We proposed a pattern-based approach for the acquisition of fined-grained NE semantic classes and instances. This corpus-based method relies only on the availability of large text corpora, such as the WaCky corpora, in contrast to resources difficult to obtain, such as query-logs (Paşca and van Durme, 2008). It makes use of a very large Web corpus to extract instances from open-domain contexts – in contrast to standard NERC approaches, which are tailored for newswire data and do not generalize well across domains. Our gold standard training and test datasets are currently based only on appositional patterns[6]. Therefore, it does not include the full spectrum of constructions in which instances can be found in context. Future work will investigate semi-supervised and heuristics (e.g. 'one sense per discourse') methods to expand the data with examples from follow-up mentions, e.g. co-occurring in the same document.

Our MaxEnt models still perform very local classification decisions, relying on separate models for each semantic class. We accordingly plan to explore both global models operating on the overall hierarchy, and more informative feature sets.

[6] The data are available for research purposes at `http://www.cl.uni-heidelberg.de/fgnerc`.

References

Enrique Alfonseca and Suresh Manandhar. 2002. An unsupervised method for general named entity recognition and automated concept discovery. In *Proc. of GWC-02*.

S. Ananiadou, C. Friedman, and J.I. Tsujii. 2004. Special issue on named entity recognition in biomedicine. *Journal of Biomedical Informatics*, 37(6).

Javier Artiles, Satoshi Sekine, and Julio Gonzalo. 2008. Web people search. In *Proc. of LREC '08*.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlap as a measure of semantic relatedness. In *Proc. of IJCAI-03*, pages 805–810.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. Maximum entropy models for named entity recognition. In *Proc. of CoNLL-03*, pages 148–151.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proc. of EACL-06*, pages 9–16.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134.

Michael Fleischman and Eduard Hovy. 2002. Fine grained classification of named entities. In *Proc. of COLING-02*, pages 1–7.

Michael Fleischman. 2001. Automated subcategorization of named entities. In *Proc. of the ACL 2001 Student Workshop*.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proc. of CoNLL-03*, pages 168–171.

Claudio Giuliano and Alfio Gliozzo. 2007. Instance based lexical entailment for ontology population. In *Proc. of ACL-07*, pages 248–256.

Claudio Giuliano and Alfio Gliozzo. 2008. Instance-based ontology population exploiting named-entity substitution. In *Proc. of COLING-ACL-08*, pages 265–272.

Claudio Giuliano. 2009. Fine-grained classification of named entities exploiting latent semantic kernels. In *Proc. of CoNLL-09*, pages 201–209.

Taku Kudo and Yuji Matsumoto. 2000. Use of Support Vector Machines for chunk identification. In *Proc. of CoNLL-00*, pages 142–144.

Yoong Keok Lee and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proc. of EMNLP-02*, pages 41–48.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the ACL-SIGDOC Conference*, pages 24–26.

Thomas Mandl and Christa Womser-Hacker. 2005. The effect of named entities on effectiveness in cross-language information retrieval evaluation. In *Proc. of ACM SAC 2005*, pages 1059–1064.

Andrew McCallum and Andrew Li. 2003. Early results for named entity recognition with conditional random fields, features induction and web-enhanced lexicons. In *Proc. of CoNLL-03*, pages 188–191.

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

MUC-7. 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Morgan Kaufmann, San Francisco, Cal.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Journal of Linguisticae Investigationes*, 30(1).

Marius Paşca and Benjamin van Durme. 2008. Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs. In *Proc. of ACL-08*, pages 19–27.

M. Paşca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. 2006a. Names and similarities on the web: Fact extraction in the fast lane. In *Proc. of COLING-ACL-06*, pages 809–816.

Marius Paşca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. 2006b. Organizing and searching the world wide web of facts – Step one: The one-million fact extraction challenge. In *Proc. of AAAI-06*, pages 1400–1405.

Marius Paşca. 2007. Weakly-supervised discovery of named entities using web search queries. In *Proc. of CIKM-2007*, pages 683–690.

Luiz Augusto Pizzato, Diego Molla, and Cécile Paris. 2006. Pseudo relevance feedback using named entities for question answering. In *Proc. of ALTW-2006*, pages 83–90.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proc. of AAAI-07*, pages 1440–1445.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A Large Ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217.

Hristo Tanev and Bernardo Magnini. 2006. Weakly supervised approaches for ontology population. In *Proc. of EACL-06*, pages 17–24.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proc. of CoNLL-03*, pages 127–132.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-independent Named Entity Recognition. In *Proc. of CoNLL-02*, pages 155–158.