THE PROSPECTS OF APPLICATION OF SEMANTIC MARKUP TO THE NAMED ENTITY RECOGNITION PROBLEM

Nekhay I. V. (nekhayiv@gmail.com)

Department of image recognition and text processing,
DIHT MIPT, Moscow, Russia

The paper describes an attempt to construct a Named Entity classifier upon ABBYY Compreno Syntactic and Semantic Parser that was presented at the "Dialogue" conference in 2012. The classifier employs supervised learning technique, namely the Conditional Random Fields model, developed under heavy constraints on the available feature set: no external NE lists or non-local features are used. The system is evaluated on the NER field's "gold standard" evaluation corpus of CoNLL-2003 achieving F-scores of 91.61% on dev and 87.51% on test set. The classifier outperforms several other systems developed under the same constraints on features, but underperforms a single system that makes use of significantly richer local context. The gain of individual classifier features based on parser attributes is explored; it is demonstrated that Compreno's semantic hierarchy and surface (syntactic) slots provide classifier with the most valuable features used to locate and classify NEs. This reliance on parser results, however, leads to error propagation from parser to classifier, as shown in the section on error analysis. Final conclusions make an attempt to offer several topics for following research.

Key words: semantic classification, named entity recognition, semantic and syntactic parser

1. Introduction

Continuing expansion of the Internet supports practical interest in methods of information extraction from unstructured texts. One subtask of information extraction is called named entity recognition (subsequently NER). This subtask consists of two problems: identification of named entity boundaries in text and further classification of named entities in a usually finite set of categories.

NER systems are usually based on text analysis systems of a much broader purpose. Analysis can vary in depth from shallow lexical or morphological, as in (Klein, Smarr, Nguyen, Manning, 2003), to integration of NER subsystem into a text parser (deep, syntactic or semantic) as described by Finkel and Manning in (Finkel, Manning, 2009). A research of capabilities of NER systems based on deep text analysis is of certain interest. Therefore a syntactic and semantic parser based on ABBYY Compreno technology, which was introduced at the "Dialog" conference in 2012 [(Anisimovich, Druzhkin, Minlos, Petrova, Selegey, Zuev, 2012) and (Bogdanov, 2012)] is the object of research described by this paper.

As the authors of the survey (Nadeau, Sekine, 2007) noted, NER solutions employ two major approaches: rule-based and statistics-based, mainly machine learning, approach. The development of a rule-based system is labour-intensive and the resulting system often trades recall for precision. The use of statistical methods, given that a sufficient amount of data is available, can significantly decrease the labour-output ratio for some tasks. That is why a statistical NER approach, using results of Compreno parser execution as source data, is considered in the current paper.

An influential comparison of language-independent NER algorithms was performed during CoNLL-2003 (Tjong Kim Sang, De Meulder, 2003) as the conference's shared task. Special corpora of news articles in English and German were prepared for this comparison. The English language CoNLL-2003 corpus has become de facto a standard for evaluation of works in the field of NER. A number of papers present results of assessments of different systems on this corpus.

Evaluation of Compreno parser applicability to the NER task in German would suffer from parser's immaturity at this moment, thus only English corpus of CoNLL-2003 was used for this research.

2. Task setup and main limitations

Evaluation method. A NER system evaluation methodology based on measurements of precision, recall and F-score was developed in the course of CoNLL conferences. The methodology will be described in greater detail in the following section. We will stick to this methodology for evaluation of our system. As our experiments show and some authors (e.g. (Tkachenko, Simanovsky, 2012), (Rosenfeld, Feldman, Fresko, 2006)) point out, the choice of features plays the most important role in the development of machine-trained NER classifiers. Therefore, the dependence of the integral F-score on the choice of features, obtained from the Compreno parser, will interest us in the first place. All our results were achieved by tuning the feature set solely; no changes or settings for a particular corpus were introduced into the learning algorithm or the parser.

Lexical features. Our approach to feature selection can be characterized as a refusal of local textual features in favour of a more generic, less language-dependent approach. The use of specific lexical features like specific separate words, word combinations, prefixes and suffixes is observed in the majority of works (CoNLL-2003 systems survey (Tjong Kim Sang, De Meulder, 2003) and later (Ratinov, Roth, 2009), (Tkachenko, Simanovsky, 2012)). In current research we don't use such features, relying on semantic descriptions of words in the semantic hierarchy, built into the parser. The resulting features are assumed to be more portable across distinct text genres, topics or even languages due to the universal inter-language nature of semantic descriptions.

External sources of information. The problem of incorporating external sources (they are also sometimes called gazetteers) such as the Wikipedia (Ratinov, Roth, 2009), DBPedia and YAGO (Tkachenko, Simanovsky, 2012) to extract list of named entities, that are later applied for NER, is another problem drawing attention of researchers. We do not use external sources, because we strive to obtain an evaluation of the parser's capabilities in «pure» form. Therefore only the published

values of F-measure which are achieved by researchers without use of external lists are chosen for comparison of results.

Local and non-local features. Due to the time constraints, we do not incorporate different types of non-local features into the system. All the explored features use only the current token, its left context, and its parent token according to the analysis tree, so the features are local only.

3. Corpus and evaluation method of CoNLL-2003

Description of the corpus. English corpus of CoNLL-2003 was created by complementing texts of Reuters news sub-corpus (about 300,000 words in size) with named entity markup according to 4 categories: person names (**PER**), organization names (**ORG**), locations (**LOC**) and all other NEs (**MISC**).

Corpus source texts were broken down into tokens, and each token has a label of respective category. CoNLL evaluation method takes into consideration the accuracy of both named entity bounds detection and classification into category. Only NEs with correctly identified bounds and category increase the values of precision, recall and F-score.

Integral F-scores, most often given in comparative papers, are calculated by micro-averaging in all four categories. More detailed description of the applied tokenization method and integral score calculation are given in (Tjong Kim Sang, De Meulder, 2003).

Corpus parts and concept drift. Initially the corpus is subdivided by its authors into three parts: training, development, and test. Training and development parts are chosen from news messages of August 1996, and test part — from reports of December 1996. As a consequence of this partition an effect, known as **concept drift**, occurs, caused by a significant change in primary persons and events appearing in news publications. A decrease of NER systems scores between development and test parts can be observed in all the papers devoted to this corpus (see table 1).

Table 1. Concept drift demonstrated by the results of top two CoNLL-2003 systems, F-score

Paper	**Florian, Ittycheriah, Jing, Zhang, 2003**			**Chieu, Ng, 2003**		
Corpus part	development	test	**drift**	development	test	**Drift**
NE label						
MISC	89.06	80.44	**−8.62**	88.41	79.16	**−9.25**
ORG	90.24	84.67	**−5.57**	88.56	84.32	**−4.24**
LOC	96.12	91.15	**−4.97**	95.57	91.12	**−4.45**
PER	96.60	93.85	**−2.75**	95.89	93.44	**−2.45**

Table 1 shows that the lowest concept drift effect can be observed in the **PER** category. It is supposedly caused by the fact that this category is the most easily formalized (and, therefore, has the most accurate markup in the corpus), and its tokens are well identifiable even by superficial lexical features, as in our study (Nekhay, 2012).

Manual exploration of classifier errors showed us that the **MISC** category, formed by residual principle, is the worst formal and contains in fact a union of other categories: nationalities, names of sports competitions, movies, etc. NEs of these categories often appear only in news messages of a narrow time period, when sufficient public interest in these subjects exists. We should also note that most errors in corpus markup ("Nato" and some other obvious NEs of category **MISC**, marked up as non-NEs) are also associated with this category.

Following assumptions can be made about the **ORG** and **LOC** categories. The usage frequency of names of important geographical objects (countries, capitals) depends at less extent on the time of publication, and name existence duration is maximal compared to **PER** and **MISC** categories. Similarly, a lot of mentions of organizations belong to large, international organizations which become international newsmakers at a constant frequency. Probably that is the reason why the **LOC** and **ORG** categories take intermediate position in table 1.

Counteracting the concept drift. In the survey (Nadeau, Sekine, 2007) its authors mention the change of text genre or domain (the latter observed between parts of CoNLL corpus) as one of the major challenges in the development of NER systems. On the other hand, in the latest works (Tkachenko, Simanovsky, 2012) the concept drift causes less significant drop of F-measure from 93.78 to 91.02. The drift is probably compensated by the use of external NE lists (gazetteers), as these lists must include with the same probability NEs appearing both in August and December of 1996. We presume that our refusal of use of local lexical features (words, suffixes, affixes, etc.) allows to reduce the effect of temporal shifts, but consider that this subject requires further research.

4. Classifier implementation

Description of the Compreno parser, upon which our classifier is based, goes beyond the current paper. The structure of the parser is described in works of (Anisimovich, Druzhkin, Minlos, Petrova, Selegey, Zuev, 2012) and (Bogdanov, 2012). As becomes clear further, the concepts of surface slot, parent and child constituent and path in the semantic hierarchy are the key concepts that form the most important features for our classifier.

4.1. Token synchronization

During the research a number of differences between data representations in the corpus markup and in the parser results were found. The most significant difference is tokenization method.

Since the parser allows adaptation of itself to a lot of different applications, we decided to keep corpus data in the original form and synchronize parse results to that form. A special algorithm that associates each corpus token to one or more parser tokens was developed for this task. However, due to the synchronization being imperfect, a certain fraction of classifier errors is caused by bad token alignment.

4.2. Classifier algorithm

Implementation. A Conditional Random Fields (CRF) model, trained by a L-BFGS algorithm provided by MALLET library (McCallum, 2002), was applied for NE identification and classification without introducing any changes to the algorithms.

Feature limitations. All features for the used algorithm are encoded as string values and treated as Booleans (true/false). Thus all the Boolean features are encoded in the most natural way — as a present feature for "true" value and as an absent feature for the opposite. Each of N-valued features is mapped into a set of N Boolean features, each corresponding to the source feature taking Nth value. This approach excludes the use of non-integer features and significantly limits available integer features to only those having a relatively small value set.

Alternative algorithms. Besides CRF, attempts were made to use decision trees and SVM for token classification. Upon increase of the number of features the main advantage of decision trees — their direct interpretation by a human — was lost, and feature mapping became rather hard to implement. SVM implementation that was applied required far more time for classifier training, though didn't provide a significant decrease of errors.

4.3. Features used by the algorithm

Before beginning to describe the features, it is important to note that the research used somewhat simplified and limited XML parser interface. This interface provides only one parse structure for each sentence (the "best" one according to the parser model) and a generalized view of the parser attributes. Attribute generalization has both positive and negative effects. While, on the one hand, it can lose significant information, it can increase classifier resilience to overfitting on the other. Access to more detailed internal parser structures was not implemented due to time limitations.

Surface-lexical features. Of the most often applied in NER field surface-lexical features, determined by token spelling, we use word case (**WCASE**: *first letter capital, all letters capital*, …) as well as a more detailed word case feature called **SHAPE**. The value of **SHAPE** is formed with a series of replacements: capital letters to the symbol "X", lowercase — to "x", digits — to "*d*". The replacements of first and last pairs of symbols remain on their places. Repetitions are excluded from other replacements, and the remaining symbols are sorted alphabetically. For example, the token "*Ireland-born*" has shape value of «*Xx-xxx*», and token «*1996-02-12*» — «*dd-ddd*». **WCASE** is considered for the current and preceding token; **SHAPE** — in the token window [−2..0].

Gazetteer-like features. External lists of NEs are not used; however, all NEs encountered during training are used to implement features named **PART_OF_(MISC|ORG|PER|LOC)**. Such feature is "true", if the current token is part of a NE in the corresponding category. To avoid overfitting these lists, randomly chosen 50% of them are used in the training phase, while at the test phase all 100% are looked through. The features are applied to the current token.

Surface-morphological features. For each word the parser determines part of speech, which is represented by our **POS** feature in [−1..0] token window.

Surface-syntactic features. For each word the parser defines two syntactic attributes: a surface slot (**SURFSL**: *Modifier_NominalEntityLike, Modifier_Attributive, Object_Indirect* …) and a simplified representation of the word's syntactic function in the sentence (**SYNTF**: *Subject, Preposition, AdverbialModifier*, …). For each token we consider these attributes of the token itself and of its parent (**PAR_SURFSL, PAR_SYNTF**), determined according the parse tree. These features are, perhaps, more dependent on the text language than others.

Deep-semantic features. The most significant for our work are features associated with semantic descriptions of words. The Compreno parser has at its foundation a semantic hierarchy (SH), which is a tree with semantic classes (SC) as nodes and lexical (roughly equivalent to word) classes (LC) as leaves. For each word the parser indicates its most probable LC and a few parent SCs along the path toward root in SH. This set of classes comprises the value of **EXLEXCLASS** feature, whose value is a vector of Booleans, corresponding to each of the parent SCs and showing, which of the SCs lie in the hierarchy path. For example, a lexical class "SOCCER" has a following set of semantic classes: *FOOTBALL : TYPES_OF_GAMES : SPORT : AREA_OF_HUMAN_ACTIVITY : … (*further parents omitted by simplified parser interface). Besides, we use several types of hierarchy path generalizations:

- Parser-defined attribute "NearestSensibleParent" (**NSP**), eliminating a lot of minor SCs. For the *soccer* example above its value would be *TYPES_OF_GAMES*.
- Artificially invented feature **ABR_EXLEXCLASS**, calculated by cutting from hierarchy path all lexical classes and semantic classes, appearing below a hardcoded list of classes, e.g. *COUNTRY_BY_NAME, PERSON_BY_FIRSTNAME* etc.
- **LEXCLASS_CONT** — a set of Boolean features, associated with the appearance in the hierarchy path of several manually selected semantic classes most correlated with NE labels in the training set.

The parser also provides an attribute named **NOUN_TYPE** that divides nouns into proper and common ones according to guesses in semantic hierarchy.

We presume that generalization of hierarchy paths plays an important role in maintaining the balance between preserving significant information and overfitting the classifier. We consider an "ideal" generalization such one that would choose for each word the most general semantic class, whose descendants in the hierarchy possess some kind of equivalence in terms of the problem being solved. However, this generalization still requires further research.

Feature combinations. Experiments demonstrate that some features (**NOUN_TYPE** and **WCASE**, **NOUN_TYPE** and **NSP**) show significantly higher results when they are combined into single feature. It is clear from intuition, that one feature with values like *(NOUN_TYPE=Common,WCASE=Lower), (NOUN_TYPE=Proper,WCASE=AllUpper)*, … possesses more information than two features valued *(Common, Proper, …)* and *(Lower, AllUpper, …)* in terms of a CRF model, based on a weighted sum of feature values. On the other hand, the values set size of a combination of several multiple-valued features may even exceed the number of words in the training set, what leads, obviously, to overfitting the data. For this reason, our classifier uses only two simple abovementioned combinations. Still, accurate feature combinations may yield higher results given enough research effort.

5. Experimental results

Table 2 shows the results achieved by our system on the CoNLL-2003 corpus and a comparison to other recent results. Results that overcome ours are highlighted with bold. It follows from the table that, with the exception of two feature sets of (Rosenfeld, Feldman, Fresko, 2006), higher results are achieved only with the help of external NE lists or document- and collection-level features.

Table 2. Comparing results to other systems

Research	Feature set	Devel. data	Test data
Our results		91.61	87.51
Tkachenko, Simanovsky, 2012	Local features	88.91	82.89
	Word + Wikipedia and DBPedia lists	85.21	78.16
	Word + Brown, Clark, LDA and phrase clusters of full Reuters corpus	90.87	87.00
	Full set (including lists)	**93.78**	**91.02**
Ratinov, Roth, 2009	(3): local features: token, word case, prefixes, suffixes, tokens in [-2..+2] window, word case in same window, two previous labels	89.25	83.65
	(3) + external lists	91.61	87.22
	(3) + Brown clusters of full Reuters corpus	90.85	86.82
	(3) + all external sources	**92.49**	**88.55**
	(3) + all non-local features	90.69	86.53
	(3) + all external + all non-local features	**93.50**	**90.57**
Rosenfeld, Feldman, Fresko, 2006	Lexical features, current and previous token		87.38
	Lexical features and combinations in [-2..0] token window		87.36
	(1): lexical features and combinations in [-2..+2] token window		**87.76**
	(2): (1) + suffixes and prefixes of previous token		**89.11**
	(2) + document- and collection-level classification results (non-local)		**90.72**
Chieu, Ng, 2003	Used training set only	91.60	86.84
	Training set and external NE lists	**93.01**	**88.31**

6. Exploring individual feature contributions

Exclusion of individual features from the common set (table 3) allows to evaluate the significance of each feature for classification. We can conclude that semantic features for current and previous token, dictionary lookup (**PART_OF**) and word case (**WCASE**, **SHAPE**) features are the most significant. It is also noticeable that semantic

features for current and previous token show a presence of correlation, meanwhile such features for the parent token are less informative.

Table 3 Impact of exclusion of individual features on the F-score; parentheses indicate area, in which the feature is calculated: t. — current token, par. — token parent, prev. — previous token

Excluded features	F- score (devel.data)	F-score (test data)
—	91.76	87.54
All semantic (t., par.)	91.41	87.23
SURFSL, SYNTF (t., par.)	91.65	87.18
NOUN_TYPE (t., prev., including all combinations)	91.45	86.97
POS (t., prev.)	91.35	86.97
All semantic (par., prev.)	90.81	86.51
WCASE (t., prev.)	91.66	86.51
SHAPE ([−2..0] window)	90.69	85.66
PART_OF (t.)	90.34	85.47
All semantic (t., prev.)	88.28	83.79
All semantic (t., par., prev.)	88.18	83.79

7. Classifier error exploration

We have explored about 100 classifier errors which show a number of widespread special cases leading to both parser and CRF-based classifier errors. Since correct parse results are statistically more common, the trained model assigns great weight to "deep" (semantic) features computed by the parser. In consequence, classifier makes a misclassification given an incorrect parse result, but some of these errors can be corrected due to presence of surface features like SHAPE and WCASE. Here are some examples of detected errors:

- Corpus text headers are given in capital letters. The parser often errs in this case; it is especially obvious when *JAPAN* receives "black varnish" semantic class and *CHINA* becomes "*porcelain*". It's evident that the parser also uses word cases to solve ambiguities.
- Name «*Bitar*» causes an error in the analysis of composites. The parser splits the name into two words which together mean "bi-resin" and the resulting features make correct classification impossible.
- In a number of cases proper names in semantic hierarchy are described in several different branches, leading to parse ambiguity. Examples of such names are *IRA* as a person name *Ira*, *Tom* as a person and a river name, *Moody* as a surname and a city name.
- Several errors made by a classifier with an absolutely correct parse tree were also found. These errors require a deeper analysis of prevalence of corresponding feature values in the training data.

8. Conclusions

We have demonstrated that the classifier built upon a CRF model and a feature set provided by the Compreno parser allows to achieve results comparable to the recent researches which do not use external NE lists and non-local features. Out of all such systems only one where a large local feature set in a wide window was considered (Rosenfeld, Feldman, Fresko, 2006) shows results that overcome ours. It lets assume that adding features based upon external NE lists and non-local document- and collection-level features to our classifier can allow reaching highest at the current moment results in NER field. However, the results of (Rosenfeld, Feldman, Fresko, 2006) also mean that greater attention should be given to features in linear and tree contexts of tokens.

Exploration of individual feature contributions shows that features related to token semantic class play the greatest role in NE identification. This fact demonstrates that universal inter-language semantic hierarchy is a rich source of information for NER solutions. Accordingly, a research of interlingual portability of a NER system based on semantic features might be of a certain interest.

Since the choice of features played an important role, in the course of the research we used different methods of automatic feature selection: individual feature exclusion, "greedy" feature inclusion, methods based on mutual information. This part is not included in the paper due to size limitations, but it is also an interesting subject for future research.

References

1. *Anisimovich K. V., Druzhkin K. J., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012). Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii «Dialog»], (pp. 90–103). Bekasovo.
2. *Bogdanov A. V.* (2012). Description of gapping in a system of automatic translation. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii «Dialog»], (pp. 61–70). Bekasovo.
3. *Chieu H. L., Ng H. T.* (2003). Named Entity Recognition with a Maximum Entropy Approach. Proceedings of CoNLL-2003, (pp. 160–163). Edmonton, Canada.
4. *Finkel J. R., Manning C. D.* (2009). Joint parsing and named entity recognition. NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, (pp. 326–334). Stroudsburg, PA, USA.
5. *Florian R., Ittycheriah A., Jing H., Zhang T.* (2003). Named Entity Recognition through Classifier Combination. Proceedings of CoNLL-2003, (pp. 168–171). Edmonton, Canada.

6. *Klein D., Smarr J., Nguyen H., Manning C. D.* (2003). Named Entity Recognition with Character-Level Models. Proceedings of CoNLL-2003, (pp. 180–183). Edmonton, Canada.
7. *McCallum A. K.* (2002). MALLET: A Machine Learning for Language Toolkit. available at: http://mallet.cs.umass.edu
8. *Nadeau D., Sekine S.* (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes , 30 (1), pp. 3–26.
9. *Nekhay I. V.* (2012). Application of n-grams and other letter- and word-level statistics to semantic classification of unknown proper nouns. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii «Dialog»], (pp. 477–489). Bekasovo.
10. *Ratinov L., Roth D.* (2009). Design challenges and misconceptions in named entity recognition. CoNLL '09 Proceedings of the Thirteenth Conference on Computational Natural Language Learning, (pp. 147–155). Stroudsburg, PA, USA.
11. *Rosenfeld B., Feldman R., Fresko M.* (2006). A Systematic Cross-Comparison of Sequence Classifiers. Proceedings of the Sixth SIAM International Conference on Data Mining. Bethesda, MD, USA: SIAM.
12. *Tjong Kim Sang E. F., De Meulder F.* (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. CONLL ,03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, 4, pp. 142–147. Stroudsburg, PA, USA.
13. *Tkachenko M., Simanovsky A.* (2012). Named entity recognition: Exploring features. KONVENS 2012, (pp. 118–127). Vienna.