Expert Systems with Applications 42 (2015) 3470–3479

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Latent semantics in Named Entity Recognition

Michal Konkol *, Tomáš Brychcín, Miloslav Konopík

Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic
NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia in Pilsen, Univerzitní 8, 306 14 Plzeň, Czech Republic

ARTICLE INFO

Article history:
Available online 20 December 2014

Keywords:
Named Entity Recognition
Information extraction
Stemming
Semantic analysis
Semantic spaces
Latent Dirichlet allocation

ABSTRACT

In this paper, we propose new features for Named Entity Recognition (NER) based on latent semantics. Furthermore, we explore the effect of unsupervised morphological information on these methods and on the NER system in general. The newly created NER system is fully language-independent thanks to the unsupervised nature of the proposed features. We evaluate the system on English, Spanish, Dutch and Czech corpora and study the difference between weakly and highly inflectional languages. Our system achieves the same or even better results than state-of-the-art language dependent systems. The proposed features proved to be very useful and are the main reason of our promising results.

1. Introduction

Named Entity Recognition (NER) systems search for important phrases (such as cities, personal names, or dates) in a given text. In this way, an NER system can serve as a valuable component for many expert systems, ranging from the standard Natural Language Processing tasks, such as question answering (Álvaro Rodrigo, Pérez-Iglesias, Peñas, Garrido, & Araujo, 2013), machine translation (Chen, Zong, & Su, 2013), social media analysis (Jung, 2012), semantic search (Habernal & Konopík, 2013), or summarization (Baralis, Cagliero, Jabeen, Fiori, & Shah, 2013; Glavaš & Šnajder, 2014; Kabadjov, Steinberger, & Steinberger, 2013) to biomedical domain (Atkinson & Bull, 2012).

The state-of-the-art NER systems are based on machine learning techniques. Many different machine learning methods have been used for NER so far. The most common examples are Hidden Markov Models (Zhou & Su, 2002), Decision Trees (Carreras, Màrquez, & Padró, 2003), Maximum Entropy (Borthwick, 1999), Support Vector Machines (Isozaki & Kazawa, 2002) and Conditional Random Fields (McCallum & Li, 2003). It has been shown that various combinations of these methods yield better results (Ekbal & Saha, 2011; Florian, Ittycheriah, Jing, & Zhang, 2003). All these methods are of the type known as supervised learning, which is the most common learning paradigm in NER. There have also been experiments with semi-supervised and unsupervised systems (Collins & Singer, 1999), but the results of such systems are significantly worse.

The NER task was defined at MUC-6 (Grishman & Sundheim, 1996). This conference was focused purely on English. The following conferences gradually attached more importance to processing multiple languages. At MUC-7/MET-2, the presented NER systems processed English, Japanese and Chinese, but it was not mandatory to evaluate the system on all these languages. In fact, the majority of the systems were evaluated on only one of these languages. For the well known CoNLL-2002 (Tjong Kim Sang, 2002) and CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), all systems had to be evaluated on a pair of languages (Dutch and Spanish, English and German). Although the systems presented at these conferences are generally considered multilingual, they had different levels of language independence. Arguably, the systems were able to adapt to a new language only to a limited extent without some expert work (e.g., part-of-speech, gazetteers were required).

In this paper, we present a machine learning based system that can be used without any change on a variety of languages with an available NE corpus and a large unlabeled corpus.

We focus on the use of semantic features. A typical semantic feature used in NER is a gazetteer (Carreras, Màrquez, & Padró, 2002; Florian et al., 2003; Konkol & Konopík, 2013), a list of named entities of the same type. Many systems use gazetteers made by human experts for a given language and domain and thus the system loses its independence to some extent. The first step in the direction of language independent semantic features were experiments with the automatic creation of gazetteers. Some approaches using both semi-supervised and unsupervised methods (Kozareva, 2006) have been published. Lin and Wu (2009) and Tkachenko and Simanovsky (2012) used word and phrase clusters, which can be seen as a substitute for a gazetteer.

* Corresponding author. Tel.: +420 377 632 491.
E-mail addresses: konkol@kiv.zcu.cz (M. Konkol), brychcin@kiv.zcu.cz (T. Brychcín), konopik@kiv.zcu.cz (M. Konopík).

http://dx.doi.org/10.1016/j.eswa.2014.12.015
0957-4174/

In this paper, we further extend this idea and exploit word similarity based on *semantic spaces* to cluster words. These clusters are then used to represent the local semantic information. We also experiment with *topic models*. They are used to represent the global semantic information. Our features are then enriched by a language-independent unsupervised stemming. We study the effects of stemming on both sources of semantic information (semantic spaces and topic models), as well as its effects on weakly and highly inflectional languages.

This research has the following goals:

- Compare our features exploiting latent semantics with other similar features.
- Explore the effects of unsupervised stemming method on both semantic features and NER system in general.
- Study the differences between various languages.

The rest of this article is organized as follows. We start with a brief introduction of latent semantics (Section 2). We follow (in Section 3) with a recapitulation of the previous work about semantic features in NER. Section 4 provides information about our NER system and the way we incorporated the novel features. Section 5 describes our experiments and also shows and discusses their results. And finally Section 6 contains our conclusions and ideas for the future work.

2. Latent semantics

In this paper, we use various methods for modeling latent semantics to improve the quality of our NER system. The basic idea behind these methods is based on distributional hypothesis (Firth, 1957) that claims *"a word is characterized by the company it keeps"*. In other words, the meaning of a word can be guessed from contexts in which it often appears. This hypothesis is supported in Rubenstein and Goodenough (1965) and Charles (2000), where authors carry out empirical tests on humans.

The computational models (that exploit this hypothesis) usually gather statistics on contexts for each word. These statistics are used to create high-dimensional vectors each representing the meaning for one word. The words represented as vectors form a vector space model. Thanks to the vector representation we can easily compare word meanings using similarities or distances of their vectors.

The methods can be roughly divided based on the context they use into *context-word* and *context-region* methods (Riordan & Jones, 2011; McNamara, 2011). In this paper, we use slightly different notation for the same division – *local context* and *global context*. A good overview of semantic models can be found in Turney and Pantel (2010), Riordan and Jones (2011) and McNamara (2011).

The *local context methods* use only a limited context around the word to infer its vector. This limited context is usually referred to as a context window and contains only a few (e.g. four) words before and after the processed word. We use the following methods for modeling the local context – HAL (Section 2.1), COALS (Section 2.2), RI (Section 2.3), BEAGLE (Section 2.4) and P&P (Section 2.5). These methods belong to a large group of algorithms known as *semantic spaces*. Later in this paper, we use the term semantic spaces as a reference to these models.

The global context methods use a much wider context, usually the whole section or document. The most prominent global context methods are LSA (Latent Semantic Analysis) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), PLSA (Probabilistic Latent Semantic Analysis) (Hofmann, 1999) and LDA (Latent Dirichlet Allocation) (Section 2.6). In this paper, we use only LDA as it represents the current state-of-the-art model for global semantics.

The local and global context methods usually discover different kinds of relations between words. For the local context approaches, the most similar words to word *hockey* can be *tennis*, *football*, or *baseball*. For the global context approaches, these can be *puck*, *player*, or *stadium*.

In the following subsections we introduce models used in this paper.

2.1. HAL

Hyperspace Analogue to Language (HAL) (Burgess & Lund, 1997; Lund & Burgess, 1996) models the similarities between words by collecting statistics about word co-occurrences. The HAL model uses two important assumptions. The first assumption is that the left context and the right context of a word contains different information and that it is important to keep their statistics separate. The second assumption is that the distance between words (in a sentence) is important and more distant words are less informative.

These assumptions are used in a creation of a co-occurrence matrix M. The size of the matrix is $|W| \times |W|$, where $|W|$ is the number of unique words in the corpus. The cell $m_{i,j}$ contains the level of co-occurrence for words w_i and w_j, more precisely for word w_j being in left context of w_i and w_i being in right context of w_j. The value $m_{i,j}$ is incremented all the times word w_j appears in the left context of w_i and the increment is weighted by the distance. If the distance between words exceeds some threshold then the word is not counted as co-occurring any more. More details about creation of the matrix can be found in Lund and Burgess (1996). Even though there is not a full information about word ordering, the model still exploits this information partially by incorporating distance weighting and side dependency of context. It is obvious that many words do not occur together so the matrix is very sparse.

The dimensionality of the matrix can be reduced using entropy. The words which are the most uniformly distributed over all other words (have the highest entropy) can be removed.

2.2. COALS

Correlated Occurrence Analogue to Lexical Semantic (or COALS) (Rohde, Gonnerman, & Plaut, 2004) is based on the combination of ideas from HAL and LSA.

The first phase of the model training is the creation of the co-occurrence matrix similarly to HAL. The difference to HAL is that it does not distinguish between left and right contexts. The co-occurrence is counted on both sides of the word and the matrix becomes symmetric. After gathering all statistics the matrix is normalized by correlation. Subsequently, all negative values are replaced by zeros and square-roots of positive values are used.

The second phase is based on LSA. Singular value decomposition is used on the matrix. This has two desired effects. The dimensionality can be rapidly reduced. The assumption is that the reduction should combine similar words together and reveal latent semantic, i.e. transitive relations between words. The second phase can be skipped for some uses.

2.3. Random Indexing

Random Indexing (RI) (Sahlgren, 2005) is based on a different approach from the previously introduced methods. The previous methods created the co-occurrence matrix from the data and context vectors were rows and columns from this matrix. The RI begins already with some initial context vectors and incrementally tries to refine them in a way that ensures similar vectors for similar contexts.