A Study of the Importance of External Knowledge in the Named Entity Recognition Task

Dominic Seyler[1], **Tatiana Dembelova**[2], **Luciano Del Corro**[2],
Johannes Hoffart[2] **and Gerhard Weikum**[2]
[1]University of Illinois at Urbana-Champaign, IL, USA
[2]Max Planck Institute for Informatics, Saarbrücken, Germany
dseyler2@illinois.edu
{tdembelo,corrogg,jhoffart,weikum}@mpi-inf.mpg.de

Abstract

In this work, we discuss the importance of external knowledge for performing Named Entity Recognition (NER). We present a novel modular framework that divides the knowledge into four categories according to the depth of knowledge they convey. Each category consists of a set of features automatically generated from different information sources, such as a knowledge-base, a list of names, or document-specific semantic annotations. Further, we show the effects on performance when incrementally adding deeper knowledge and discuss effectiveness/efficiency trade-offs.

1 Introduction

Named Entity Recognition (NER) is the task of detecting named entity mentions in text and assigning them to their corresponding type. It is a crucial component in a wide range of natural language understanding tasks, such as named entity disambiguation (NED), question answering, etc.

Previous work (Ratinov and Roth, 2009) argued that NER is a knowledge-intensive task and used prior knowledge with outstanding results. In this work, we attempt to quantify to which extent external knowledge influences NER performance. Even though recent approaches have excelled in end-to-end neural methods, this paper aims to give transparency and user-comprehensible explainability. This is especially significant for industrial sectors (e.g., those heavily regulated) that require the use of transparent methods for which a particular decision is explainable.

We perform the study by devising a simple modular framework to exploit different sources of external knowledge. We divide the information sources into four different categories according to the depth of knowledge they convey, each one carrying more information than the previous. Each category is composed of a set of features that reflect the degree of knowledge contained in each source. Then, we feed a linear chain CRF, a transparent, widely used method used for NER.

We perform our experiments on two standard datasets by testing various combinations of knowledge categories. Our results indicate that the amount of knowledge is highly correlated with NER performance. The configurations with more external knowledge systematically outperform the more agnostic ones.

2 Knowledge Augmented NER

In the following section, we describe the four knowledge categories in detail. Table 1 gives an overview of the features on the categories that use external knowledge. The features were used to train a linear chain CRF, a simple and explainable method, proven to work well for NER (Finkel et al., 2005; Jun'ichi and Torisawa, 2007; Ratinov and Roth, 2009; Passos et al., 2014; Radford et al., 2015).

2.1 Knowledge Agnostic (A)

This category contains the "local" features, which can be extracted directly from text without any external knowledge. They are mostly of a lexical, syntactic or linguistic nature and have been well-studied in literature. We implement most of the features described in (Finkel et al., 2005):

(1) The current word and words in a window of size 2; (2) Word shapes of the current word and words in a window of size 2; (3) POS tags in a window of size 2; (4) Prefixes (length three and four) and Suffixes (length one to four); (5) Presence of the current word in a window of size 4; (6)

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 241–246
Melbourne, Australia, July 15 - 20, 2018.

Cat.	Feature	Description	Example
Name	**Mention tokens**	Some tokens are strongly associated to NEs	county,john,school,station,...
	POS-tag sequence	Multi-word NEs tend to share POS patterns	Organization of American States → NNP IN NNP NNP
KB	Type gazetteers	Names that are associated to types	Florida → location
	Wiki. link prob.	Tokens that are associated to NEs	"Florida" linked in Wikipedia
	Type prob.	Probability of token to type associations	Obama → person;
Entity	Doc. gazetteers	NE presence indicates other NEs	European Union → EU

Table 1: Features by category (novel features are highlighted)

Beginning of sentence.

2.2 Name-Based Knowledge (Name)

Here, the knowledge is extracted from a list of named entity names. These features attempt to identify patterns in names and exploit the fact that the set of distinct names is limited. We extracted a total of more than 20 million names from YAGO (Suchanek et al., 2007) and derived the following features:

Frequent mention tokens. Reflects the frequency of a given token in a list of entity names. We tokenized the list and computed frequencies. The feature assigns a weight to each token in the text corresponding to their normalized frequency. High weights should be assigned to tokens that indicate named entities. For instance, the top-5 tokens we found in English were "county", "john", "school", "station" and "district". All tokens without occurrences are assigned 0 weight.

Frequent POS Tag Sequences. Intends to identify POS sequences common to named entities. For example, person names tend to be described as a series of proper nouns, while organizations may have richer patterns. Both "Organization of American States" and "Union for Ethical Biotrade" share the pattern NNP-IN-NNP-NNP. We ranked the name POS tag sequences and kept the top 100. The feature is implemented by finding the longest matching sequences in the input text and marking whether the current token belongs to a frequent sequence or not.

2.3 Knowledge-Base-Based Knowledge (KB)

This category groups features extracted from a KB or an entity annotated corpus. They encode knowledge about named entities themselves or their usages. We implemented three features:

Type-infused Gazetteer Match. Finds the longest occurring token sequence in a type-specific gazetteer. It adds a binary indicator to each token, depending on whether the token is part of a sequence. We use 30 dictionaries distributed by (Ratinov and Roth, 2009) containing type-name information for English. These dictionaries can also be created automatically by mapping each dictionary to a set of KB types and extracting the corresponding names. This automatic generation is useful in multilingual settings, which we discuss in Section 3.5.

Wikipedia Link Probability. This feature measures the likelihood of a token being linked to a named entity Wikipedia page. The intuition is that tokens linked to named entity pages tend to be indicative of named entities. For instance, the token "Obama" is usually linked while "box" is not. The list of pages referring to named entities is extracted from YAGO. Given a token in the text, it is assigned the probability of being linked according to Eq. 1, where $link_d(t)$ equals 1, if token t in document d is linked to another Wikipedia document. $present_d$ equals 1 if t occurs in d.

$$P_{Wiki}(t) = \frac{\sum_{d \in D} link_d(t)}{\sum_{d \in D} present_d(t)} \quad (1)$$

Type Probability. Encodes the likelihood of a token belonging to a given type. It captures the idea that, for instance, the token "Obama" is more likely a person than a location. Given a set of entities E in YAGO with mentions M_e and tokens T_{em} we calculate the probability of a class $c \in C$ given a token t as in Eq. 2, where $c(e) = 1$ if entity e belongs to class c and $c(e) = 0$ otherwise. For each token in the text, we create one feature per type with the respective probability as its value.

$$P(c|t) = \frac{\sum_e^E \sum_{m_e}^{M_e} \sum_{t_{em}}^{T_{em}} c(e)}{\sum_e^E \sum_{m_e}^{M_e} \sum_{t_{em}}^{T_{em}} \sum_{c_i}^{C} c_i(e)} \quad (2)$$

Token Type Position. Reflects that tokens may appear in different positions according to the entity type. For instance, "Supreme Court of the United States", is an organization and "United"

occurs at the end. In "United States", a location, it occurs at the beginning. This helps with nested named entities.

This is implemented using the BILOU (**B**egin, **I**nside, **L**ast, **O**utside, **U**nit) encoding (Ratinov and Roth, 2009), which tags each token with respect to the position in which it occurs. The number of features depends on the number of types in the dataset (4 BILU positions times n classes + O position). For each token, each feature receives the probability of a class given the token and position. The class probabilities are calculated as in Equation 2, incorporating also the token position.

As a result, for each token we now have a probability distribution over $4n + 1$ classes. Take for instance the token "Obama". We would expect it to have high probability for classes "B-Person" (i.e., last name in combination with first name) and "U-Person" (i.e., last name without first name). The probabilities for all other classes would be close to zero. In comparison, the word "box" should have high probability for class "O" and close to zero for all others, since we would not expect it to occur in many named entities.

2.4 Entity-Based Knowledge (Entity)

This category encodes document-specific knowledge about the entities found in text to exploit the association between NER and NED. Previous work showed that the flow of information between these generates significant performance improvements (Radford et al., 2015; Luo et al., 2015).

Comparatively, this module needs significantly more computational resources. It requires a first run of NED to generate document specific features, based on the disambiguated named entities. These features are used in a second run of NER.

Following (Radford et al., 2015), after the first run of NED, we create a set of document-specific gazetteers derived from the disambiguated entities. This information helps in the second round to find new named entities that were previously missed. Take the sentence "Some citizens of the European Union working in the United Kingdom do not meet visa requirements for non-EU workers after the uk leaves the bloc". We can imagine that in the first round of NED *European Union* and *United Kingdom* can be easily identified but "EU" or the wrongly capitalized "uk" might be missed. After the disambiguation, we know that both entities are organizations and have the aliases *EU* and *UK* respectively. Then, in a second round it may be easier to spot mentions "EU" and "uk".

After a first run of NER+NED, we extract all surface forms of the identified entities from YAGO. These are tokenized and assigned the type of the corresponding entity plus its BILOU position. For example, the surface form "Barack Obama" results in "Barack" and "Obama", assigned to "B-Person" and "L-Person". There are 17 binary features (BILU tags multiplied by four coarse-grained types + O tag), which fire when a token is part of a list that contains the mappings from tokens to type-BILOU pairs.

3 Evaluation

3.1 Experimental Setup

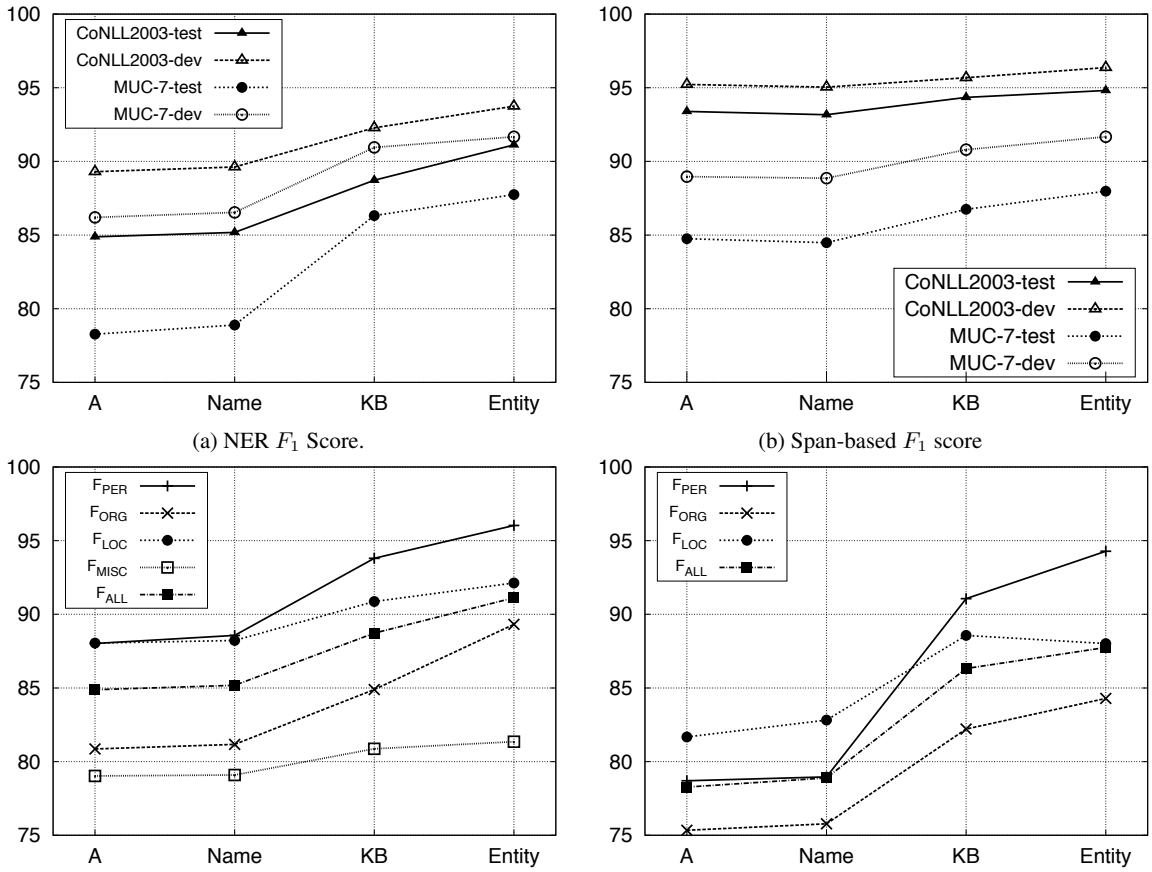
System Setup. To perform our study we use a linear chain CRF(Lafferty et al., 2001). CRFs are transparent and widely used for NER (Finkel et al., 2005; Jun'ichi and Torisawa, 2007; Ratinov and Roth, 2009; Passos et al., 2014; Radford et al., 2015; Luo et al., 2015). The entity-based component was implemented using the AIDA (Hoffart et al., 2011) entity disambiguation system.

Datasets. We evaluate on two standard NER datasets *CoNLL2003*. (Sang and Meulder, 2003), a collection of English newswires covering entities with four types (PER, ORG, LOC, MISC) and *MUC-7*, a set of New York Times articles (Chinchor and Robinson, 1997) with annotations on three types of entities (PER, ORG, LOC).

3.2 Incremental knowledge

Here we analyze the impact of incrementally adding external knowledge. Fig. 1a shows four variants. Each contains the features corresponding to a given category plus all those from the lighter categories to the left. In all cases adding knowledge boosts F_1 performance. The effect is particularly strong for MUC-7-test which registered an overall increment of almost 10 points. In both datasets, the biggest boost is registered when the KB-based features are added. As a reference point, one of the best systems to date (Chiu and Nichols, 2016) (neural-based) achieves F_1 91.62 on CoNLL2013-test, while our full-knowledge CRF reaches F_1 91.12.

Fig. 1c shows the performance for each entity type on CoNLL2003. Again, there is a boost in all cases, especially organizations. Persons also improve significantly: At first they perform similar to



(a) NER F_1 Score.

(b) Span-based F_1 score

(c) Type-based NER F_1 score on CoNLL2003-test

(d) Type-based NER F_1 score on MUC-7-test

Figure 1: Evaluation results by type CoNLL2003-test and MUC-7-test.

locations, but the successive increment is sharper. F_1 for persons achieves F_1 96.03 and locations F_1 92.13. The positive effect is quite significant for organizations (F_1 80.86 to F_1 89.32), while it is moderate for miscellaneous. Fig. 1d shows results by type for MUC-7-test. The positive effect is particularly strong for persons, improving more than 15 F_1 points (78.70 to 94.28). Interestingly, locations register a slight decline between KB and Entity (0.56 F_1 points).

Finally, Fig. 1b shows the performance over span detection, which is the span where the named entity occurs without taking type information into account. This is especially important for applications such as named entity disambiguation. It drops slightly for the name-based category, but it increases again as more knowledge is added. The effect is similar on both datasets.

3.3 Ablation

Table 2 shows different combinations of knowledge categories. The relatively small improvement from KB to Entity suggests that KB features are subsumed by the later. This is somehow expected as the entity specific information is extracted from the same KB and both rely on entity types. However, as we will see, this comes at a cost.

Feature Categories	F_1
A, Name, KB	88.73
A, Name, Entity	89.32
A, KB, Entity	91.09
All	91.12

Table 2: : Ablation study by categories on CoNLL2003-test

3.4 Timing

The Entity-based component is by far the most expensive concerning timing performance. We measure 314ms, 494ms, 693ms, and 4139ms for A, Name, KB and Entity based features, respectively (Figure 2). Since KB-based features are

comparable in performance to the Entity-based features, but the latter are much more expensive, these findings allow practitioners to carefully decide whether the additional computational cost is worth the relatively small performance improvements. The modularity of our feature classes allows for optimal tuning of a system regarding effectiveness/efficiency trade-offs.

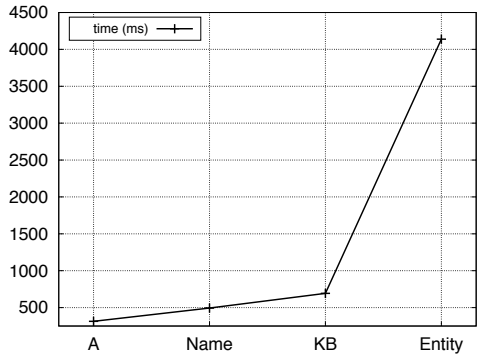


Figure 2: Timing experiments for CoNLL2003e-test in average milliseconds per document

3.5 Multilingualism

In order to demonstrate the general applicability of our approach, we implement our NER system for two additional languages, namely German and Spanish. All features for the Name, KB and Entity knowledge classes are derived from the respective language's Wikipedia. Performance is evaluated on CoNLL2003g (Sang and Meulder, 2003) for German and CoNLL2002 (Tjong Kim Sang, 2002) for Spanish. Results can be found in Figure 3. Similar to the performance on English data, we can see that adding more external knowledge improves performance. For reference, we found that performance is close to the state-of-the art in both languages. Our system lags only 1.56 F_1 points on (Lample et al., 2016) in German and 1.98 F_1 points on (Yang et al., 2016) in Spanish.

4 Related Work

NER is a widely studied problem. Most of previous work rely on the use of CRFs (Finkel et al., 2005; Jun'ichi and Torisawa, 2007; Ratinov and Roth, 2009; Passos et al., 2014; Radford et al., 2015; Luo et al., 2015). A recent trend has achieved particularly good results modeling NER as an end-to-end task using neural networks (dos Santos and Guimarães, 2015; Chiu and Nichols, 2016; Lample et al., 2016; Yang et al., 2016; Gillick et al., 2016). While this constitutes a big step forward, certain applications (e.g., in heavily regulated sectors) require a degree of explainability that neural approaches cannot yet provide.

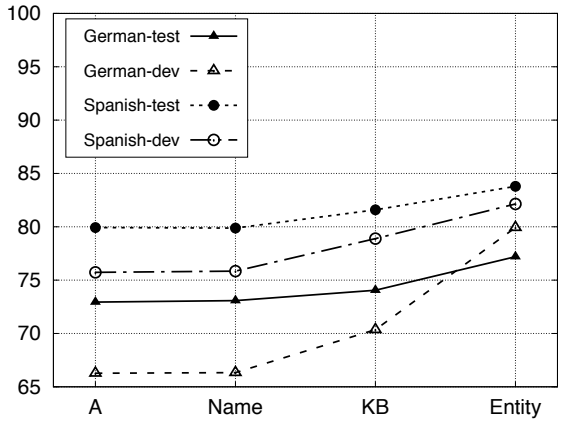


Figure 3: NER F_1 for German on CoNLL2003g dataset and Spanish on CoNLL2002 dataset.

Previous work has already regarded NER as a knowledge intensive task (Florian et al., 2003; Zhang and Johnson, 2003; Jun'ichi and Torisawa, 2007; Ratinov and Roth, 2009; Lin and Wu, 2009; Passos et al., 2014; Radford et al., 2015; Luo et al., 2015). Most of these works incorporate background knowledge in the form of entity-type gazetteers (Florian et al., 2003; Zhang and Johnson, 2003; Jun'ichi and Torisawa, 2007; Ratinov and Roth, 2009; Passos et al., 2014). Others, used external knowledge by exploiting the association between NER and NED (Durrett and Klein, 2014; Radford et al., 2015; Luo et al., 2015; Nguyen et al., 2016). In this study, we attempt to bring more light on the issue by quantifying the effect of different degrees of external knowledge. Our modular framework allows to test this intuition via novel feature sets that reflect the degree of knowledge contained in available knowledge sources.

5 Conclusion

We investigated the importance of external knowledge for performing Named Entity Recognition by defining four feature categories, each of which conveys a different amount of knowledge. In addition to commonly used features in existing literature, we defined four novel features that we incorporated into our category scheme. We experimentally showed that although more external knowledge leads to performance improvements, it comes at a considerable performance trade-off.

References

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of MUC-7*.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns.

Cícero Nogueira dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. *Proceedings of the Fifth Named Entity Workshop* .

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. In *TACL*.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.

Radu Florian, Abraham Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of CoNLL*.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of NAACL*.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of EMNLP*.

Kazama Jun'ichi and Kentaro Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of EMNLP-CoNLL*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL*.

Dekang Lin and Xiaoyun Wu. 2009. Phrase clustering for discriminative learning. In *Proceedings of ACL*.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of EMNLP*.

Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2016. J-NERD: joint named entity recognition and disambiguation with rich linguistic features. *TACL* 4:215–229.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of CoNLL*.

Will Radford, Xavier Carreras, and James Henderson. 2015. Named entity recognition with document-specific KB tag gazetteers. In *Proceedings of EMNLP*.

Lev-Arie Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of WWW*.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR* .

Tong Zhang and David Johnson. 2003. A robust risk minimization based named entity recognition system. In *Proceedings of CoNLL*.