Named entity recognition with document-specific KB tag gazetteers

Will Radford **Xavier Carreras** **James Henderson**
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
firstname.lastname@xrce.xerox.com

Abstract

We consider a novel setting for Named Entity Recognition (NER) where we have access to *document-specific knowledge base tags*. These tags consist of a canonical name from a knowledge base (KB) and entity type, but are not aligned to the text. We explore how to use KB tags to create document-specific gazetteers at inference time to improve NER. We find that this kind of supervision helps recognise organisations more than standard wide-coverage gazetteers. Moreover, augmenting document-specific gazetteers with KB information lets users specify fewer tags for the same performance, reducing cost.

1 Introduction

NER is the task of identifying names in text and assigning them a type (e.g. person, location, organisation, miscellaneous). State-of-the-art supervised approaches use models that incorporate a name's form, its linguistic context and its compatibility with known names. These models rely on large manually-annotated corpora, specifying name spans and types. These are vital for training models, but it is laborious and expensive to label every occurrence of a name in a document.

We consider a non-standard setting where, for each document, we have metadata in the form of *document-specific knowledge base tags*. A KB tag is a canonical name, that is an identifier in a KB (e.g. a Wikipedia title), and an entity type. While these tags have a correct type assigned for at least one context, they are not aligned to phrases in the text, and may not share the same form as all of their mentions (e.g. we may see the tag `United Nations` for the mention UN). We also assume that each tag matches at least one mention in the document, but do not specify where in the document the mention is.

There are many sources of KB tags, such as manual entity indexing for news stories or data extracted from personalised knowledge stores. For example, the New York Times Annotated Corpus (Sandhaus, 2008) contains more than 1.5M articles "manually tagged by library scientists with tags drawn from a normalized indexing vocabulary of people, organizations, locations and topic descriptors". Names and types are also present in large quantities of financial news stories from Bloomberg (Bradesko et al., 2015), in the form of linked names of companies and people.

Document-level tags may be quicker for annotators to apply than the usual method of marking spans in text, and are thus a cheap form of supervision. It is hard to make strong comparisons to the standard NER task, as KB tags can be considered partial, unaligned gold-standard supervision – so fully supervised models *should* perform better, the question is by how much and why.

This paper explores effective ways to use KB tags for improving NER. We use the CoNLL 2003 English NER dataset (Tjong Kim Sang and De Meulder, 2003), annotated with Wikipedia links (Hoffart et al., 2011). This allows us to simulate a set of KB tags for each document in the TRAIN, TESTA and TESTB splits of the dataset. We use a document's KB tags to build a document-specific gazetteers which we use in addition to standard features for a conditional random field (CRF) model (Lafferty et al., 2001).

We compare against wide-coverage gazetteers, which score 89.85% F-score on TESTA. Assuming access to all possible KB tags, the upper bound for KB tag models is substantially better at 92.85% F-score. KB tags help NER accuracy across all entity types, but provide relatively better supervision for organisation entities than wide-coverage gazetteers. The benefit of KB tags comes from their type information, which is required for good performance. We also examine how performance

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 512–517,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

degrades as we use fewer KB tags, simulating the use-case where a busy knowledge worker spends less time annotating. We find that KB augmentation means we require fewer tags to reach the same performance, which reduces the cost of obtaining KB tags. We show how KB tags can be exploited as a useful complement to traditional NER supervision.

2 Background

Gazetteers have long been used to augment statistical NER models, adding general evidence of tokens used in names (Nadeau and Sekine, 2007). These are usually drawn from wide-coverage sources like Wikipedia and census lists (Ratinov and Roth, 2009) and can be incorporated into sequence models by designing binary features that indicate whether a token appears in a gazetteer entry. Features can be refined by specifying which part of an entry a token matches using tag encoding schemes such as IOB (Kazama and Torisawa, 2007). Using multiple gazetteers allows feature weights to capture different name types and sources. Given their purpose to increase coverage beyond names included in training data, gazetteers are usually large, general and static, remaining the same during training and prediction time.

Beyond their use as sources for gazetteers, the link structure in and around KBs has been used to create training data. A prominent technique is to follow links back from KB articles to documents that mention the subject of the article, heuristically labelling high-precision matches to create training data. This has been used for genetic KBs (Morgan et al., 2003; Vlachos and Gasperin, 2006), and Wikipedia (Kazama and Torisawa, 2007; Richman and Schone, 2008; Nothman et al., 2013). These works do not consider our setting where gold-standard entities are given at inference time as their goal is to generate training data.

KBs have also been used to help other natural language processing tasks such as coreference resolution (Rahman and Ng, 2011), topic modelling (Kataria et al., 2011) and named entity linking (Cucerzan, 2007; Ratinov et al., 2011). Finally, it may be that supervised data is only available in some circumstances, for example in the case of personalising NER models. Jung et al. (2015) query a user's smartphone data services to create user-specific gazetteers of personal information. The background NER model is initially trained without access to the user-specific information and later adapted on the users's smartphone.

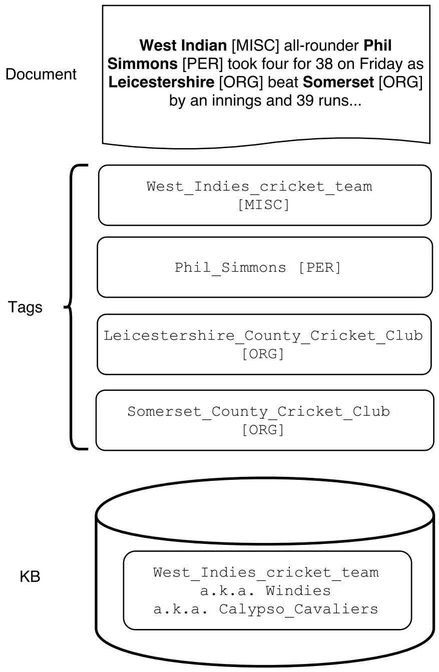
Document
West Indian [MISC] all-rounder **Phil Simmons** [PER] took four for 38 on Friday as **Leicestershire** [ORG] beat **Somerset** [ORG] by an innings and 39 runs...
Tags
West_Indies_cricket_team [MISC]
Phil_Simmons [PER]
Leicestershire_County_Cricket_Club [ORG]
Somerset_County_Cricket_Club [ORG]
KB
West_Indies_cricket_team a.k.a. Windies a.k.a. Calypso_Cavaliers

Figure 1: An entity-tagged document, KB tags with canonical name and type and KB with aliases.

3 Document-level KB tags

We incorporate information from KB tags by building document-specific gazetteers. Figure 1 shows an example with a document in which the **names** need to be recognised and typed (in square brackets). We are also given a list of KB tags, each of which is a canonical name and a type. These are linked to a KB which we use to extract aliases, in our case each canonical name is a Wikipedia article, and redirects to that article are considered aliases. Our goal is that knowing that a document mentions the entity `West Indies cricket team` can help us identify Windies or Calypso Cavaliers.

To create gazetteers from a document's KB tags, we preprocess the canonical name from each KB tag, tokenising by underscore, lowercasing and removing parenthesised suffixes (e.g. `Chris_Lewis_(cricketer)` becomes chris lewis). We use an encoding scheme to incorporate the type information from the KB tag. Inspired by Kazama and Torisawa (2007), who applied IOB encoding to gazetteers, we apply the BMEOW (a.k.a. BILOU), a scheme that also distinguishes between **b**eginning, **m**iddle, **e**nd, **o**utside and single word positions.[1] For example, this allows us to map chris lewis to B-PER E-PER, and we can aggregate gazetteers of tokens for each encoded type, such that the gazetteer for B-PER contains chris.

Our CRF gazetteer features are calculated from an input token from the text that we wish to label. Having created a document's KB tag gazetteers, we can define binary features that are active if an input token matches (case-insensitively) with a particular gazetteer. This models both the part of the KB tag name that the token matched, and its type. The input token Chris thus activates the feature fB-PER and the token cricket would activate the fI-MISC and fI-ORG, as it matches inside entries of the two types.

4 Methodology

We define several configurations to investigate KB tags. The first four baselines either do not use KB tags, or do not integrate them into the CRF. The second four configurations use KB tag features in the CRF model.

4.1 Baselines

KB tag matching (MATCH) We find the longest full match from the document gazetteer and apply the known type. This will not match partial or non-canonical names, but should be high-precision. This is similar to the CoNLL 2003 baseline system (Tjong Kim Sang and De Meulder, 2003).

Baseline (CRF) We train a CRF model using CRFsuite (Okazaki, 2007) with a standard set of features that encode lexical context, token shape, but no external knowledge features such as gazetteers. All following configurations build on the CRF with standard features.

KB tag repair (CRF$_{+\text{REPAIR}}$) We label the text using the baseline CRF, then find the longest full match from the document gazetteer and assign the known type. When a gazetteer match overlaps with a CRF match, we prefer the gazetteer and remove the latter. Although we do not consider partial matches, this may recognise longer names that can be difficult for CRF models.

Wide-coverage gazetteers (CRF$_{+\text{WIDE}}$) This uses gazetteers distributed with the Illinois NER system (Ratinov and Roth, 2009). We encode each phrase using the BMEOW scheme described above, and use the filename of each gazetteer as its type. There are 33 gazetteers drawn from many sources with approximately 2 million entries.

4.2 Using KB tags as CRF features

KB tag names (CRF$_{+\text{NAME}}$) We generate document-specific gazetteer features, but use the same type for each entry.

KB tag names and types (CRF$_{+\text{NAME+TYPE}}$) This is equivalent to CRF$_{+\text{NAME}}$, but includes known types. Since type varies with context, this may not be correct, but is hopefully informative.

KB tag names, types and KB aliases (CRF$_{+\text{NAME+TYPE+AKA}}$) This builds on the above, but uses the KB to augment the document-specific gazetteer with known aliases of the KB tags, for example adding UN for United_Nations with the known type.

KB tag names, types, KB aliases and large gazetteers (CRF$_{+\text{NAME+TYPE+AKA+WIDE}}$) This combines all KB tag features with the wide-coverage gazetteers.

We fetch and cache KB information using a Wikipedia API client.[2] We assume the tag set of person (PER), organisation (ORG), location (LOC) and miscellaneous (MISC), and report precision, recall and F-score from the conlleval evaluation script. The median proportion of mentions in a document that are linked to the KB is 81% in TRAIN and TESTB, and 85% in TESTA. Augmenting the gazetteer with aliases produces, on average, 26 times the number of gazetteer entries than KB tags alone in TESTA, and 23 times in TESTB.

5 Results

Table 1 shows the performance of different configurations – we focus first on TESTA overall F-scores. Matching against KB tag names results in high-precision but low recall with an F-score of 55.35%, far worse than the baseline CRF at 87.68%. Despite its naïve assumptions, repairing the CRF tags using longest matches in the document gazetteer performs surprisingly well at 89.76%, just lower than using wide coverage gazetteers, with an F-score of 89.85%.

The first setting that uses KB tags as CRF features is CRF$_{+\text{NAME}}$, which includes typeless names

[1] We omit the O tag as all gazetteer tokens are inside.

[2] https://github.com/goldsmith/Wikipedia adapted to allow access to Redirect pages.

Method	TESTA							TESTB		
	P	R	F	F_{LOC}	F_{MSC}	F_{ORG}	F_{PER}	P	R	F
MATCH	94.93	39.06	55.35	76.90	22.01	29.47	59.24	94.57	37.62	53.83
CRF	88.52	86.86	87.68	90.92	85.18	81.44	90.06	81.87	80.93	81.40
+REPAIR	89.28	90.24	89.76	91.68	87.44	84.44	92.68	84.06	86.54	85.28
+WIDE	90.45	89.26	89.85	92.63	85.99	84.21	93.00	85.10	83.80	84.44
+NAME	89.96	88.64	89.29	92.21	85.57	82.71	92.89	84.26	82.72	83.48
+NAME+TYPE	93.29	92.12	92.70	95.42	88.19	88.18	95.38	89.46	87.92	88.69
+NAME+TYPE+AKA	93.42	92.29	92.85	95.63	88.35	88.27	95.54	89.90	88.74	89.32
+NAME+TYPE+AKA+WIDE	93.13	92.01	92.57	95.48	88.34	87.96	94.97	89.86	88.85	89.35

Table 1: Results for CoNLL 2003 TESTA and TESTB. We report P/R/F for all tags and per-type F-scores. Methods starting with "+" build on the standard CRF by repairing or adding features.

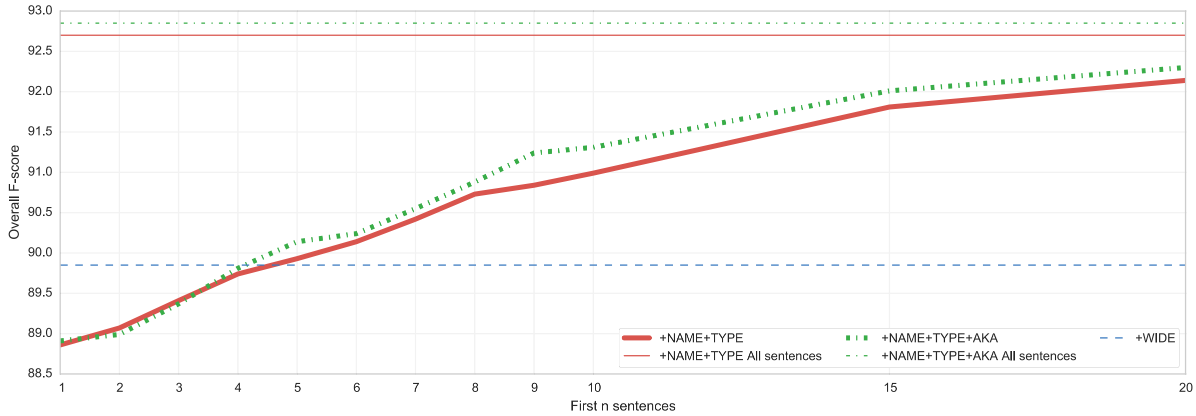

Figure 2: How many sentences should an annotator check for KB tags? TESTA results for CRF$_{+\text{NAME+TYPE}}$ and CRF$_{+\text{NAME+TYPE+AKA}}$ where KB tags are drawn from the first n sentences. This is compared to the F-score for CRF$_{+\text{WIDE}}$ and versions of the models with access to all sentences in the document (horizontal, thin lines).

and has an F-score of 89.29%. Precision and recall are lower than wide coverage gazetteers, suggesting that, without type information, bigger gazetteers are better. Adding type features (CRF$_{+\text{NAME+TYPE}}$) results in better performance than either CRF or CRF$_{+\text{WIDE}}$ at 92.7% F-score.[3] Augmenting the document gazetteers using aliases from the KB further improves F-score for aliases (92.85%). Adding wide-coverage gazetteers to KB tags slightly decreases F-score at 92.57%. These results indicate that type information is critical and, to confirm this, we ran experiments that used only name and alias information from KB tags. This scores 89.45% F-score on TESTA and 83.62% F-score on TESTB. While aliases help, type information is required to improve performance beyond wide coverage gazetteers.

To give some insight into why KB tag types are effective, consider the name West Indian. This appears 65 times across 11 of the 33 wide-coverage gazetteers, including those that also contain people, locations, organisations, songs. A document-specific gazetteer is able to constrain this type ambiguity, producing a cleaner signal for the model. Aliases, on the other hand, allow the model to capture non-canonical variants of a name, but this depends on type information for good NER performance.

We also examine the per-tag F-scores for TESTA to investigate whether KB tags help some types of entities more than others. Using CRF$_{+\text{NAME+TYPE+AKA}}$ we obtain around 95.5% F-score for PER and LOC entities. As with CRF$_{+\text{WIDE}}$, MISC entities remain hard to tag correctly. However, if we consider the percentage F-score gain by type over the CRF baseline, CRF$_{+\text{WIDE}}$ gazetteers improve performance most for PER (+2.94%), then ORG (2.77%) entities. The top two are reversed for CRF$_{+\text{NAME+TYPE+AKA}}$, with ORG (+6.83%), then PER (+5.48%). This suggests that KB tags are particularly well-suited for helping recognise organisations names.

[3] We tried to manually map the 33 gazetteer filenames to the 4 NER types, but this reduced performance on TESTA.

We see similar trends in TESTB, except KB tags and CRF$_{\text{+WIDE}}$ are complementary. The experiments illustrate that if we are lucky enough to have KB tags, they improve NER. However, the models use all possible KB tags and should be considered an upper bound. To better model busy workers, we restrict the gazetteers to only KB tags from mentions in the first n sentences. This matches asking an annotator to only bother looking at the first n sentences. Figure 2 shows how the KB tag models perform on TESTA as we increase n. To achieve better performance than CRF$_{\text{+WIDE}}$, one should view the first 5 sentences for CRF$_{\text{+NAME+TYPE}}$. Aliases (CRF$_{\text{+NAME+TYPE+AKA}}$) reduce performance slightly when only using a few sentences, but with more than 4 sentences, aliases are consistently useful. This trend is also apparent in TESTB, showing that augmenting tags with KB information improves NER, especially when only a few tags are available.

6 Discussion and conclusion

There are several avenues to explore further. Asking annotators to specify types is not ideal and it would be better to predict them from the KB. We only use the KB to collect aliases, but we could use it to harvest related entities. Another challenge is appropriately modelling the interaction between a sentence-level task and document-level constraints. A KB tag might match a mention in one sentence and this should influence predictions there. However, its evidence should be less important elsewhere since that constraint has already been satisfied. This would improve robustness, however global constraints are hard to model in sentence-unit CRF models.

This paper presents a novel NER setting whereby we have access to some number of KB tags – canonical names and types – at training and inference time. We explore how best to use this information, finding that CRF models can indeed take advantage of this non-standard supervision. Moreover, models benefit from integration with the KB, in our case augmenting document gazetteers to maximise the benefit of KB tags.

Acknowledgements

Thank you to the anonymous reviewers whose comments helped us improve the paper.

References

Luka Bradesko, Janez Starc, and Stefano Pacifico. 2015. Isaac Bloomberg Meets Michael Bloomberg: Better Entity Disambiguation for the News. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 631–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June. Association for Computational Linguistics.

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 the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

YoungHoon Jung, Karl Stratos, and Luca P. Carloni. 2015. LN-Annote: An alternative approach to information extraction from emails using locally-customized named-entity recognition. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 538–548, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Saurabh S Kataria, Krishnan S Kumar, Rajeev R Rastogi, Prithviraj Sen, and Srinivasan H Sengamedu. 2011. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1037–1045. ACM.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707, Prague, Czech Republic, June. Association for Computational Linguistics.

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 the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alex Morgan, Lynette Hirschman, Alexander Yeh, and Marc Colosimo. 2003. Gene name extraction using flybase resources. In *Proceedings of the ACL*

2003 Workshop on Natural Language Processing in Biomedicine, pages 1–8, Sapporo, Japan, July. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January. Publisher: John Benjamins Publishing Company.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194(0):151 – 175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 814–824, Portland, Oregon, USA, June. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado, June. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of ACL-08: HLT*, pages 1–9, Columbus, Ohio, June. Association for Computational Linguistics.

E. Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Andreas Vlachos and Caroline Gasperin. 2006. Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 138–145, New York, New York, June. Association for Computational Linguistics.