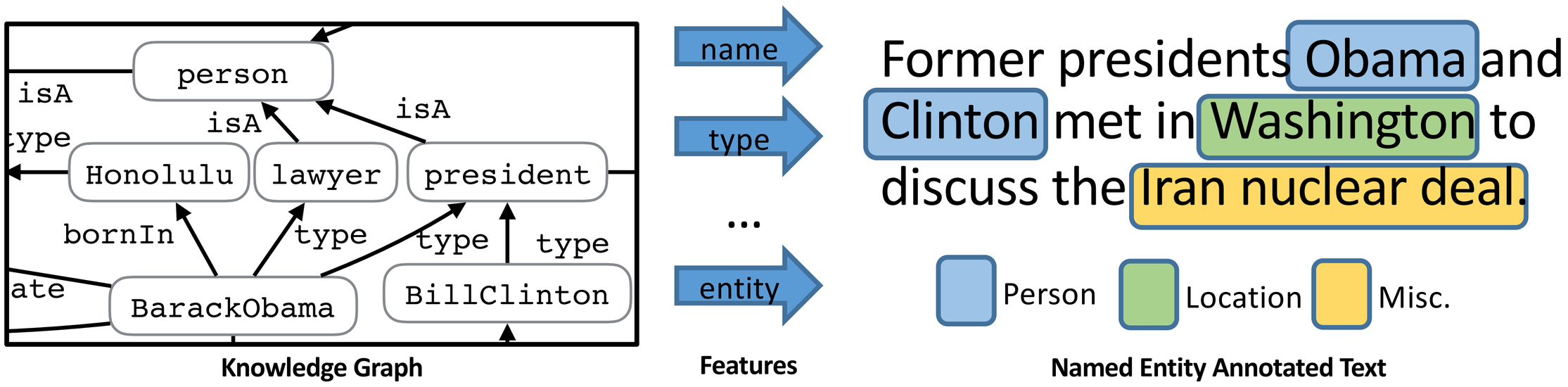
A Study of the Importance of External Knowledge in the Named Entity Recognition Task

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

Overview



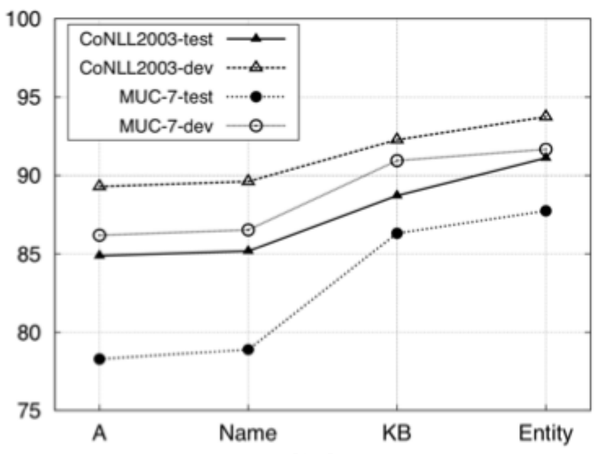
Features

Knowledge Augmented NER Modules

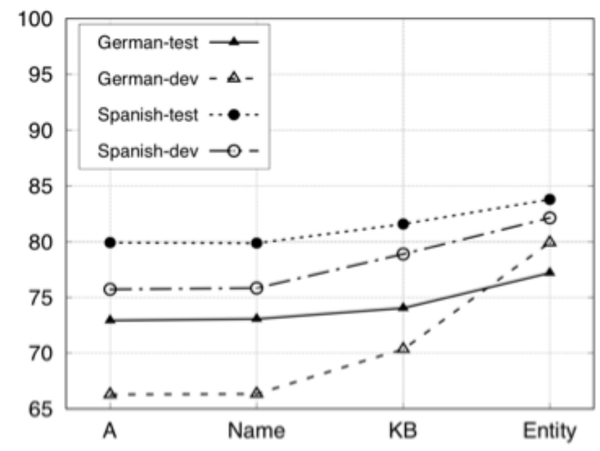
1. **Knowledge Agnostic (A):** "local" features, which can be extracted directly from text without any external knowledge.
2. **Name-Based Knowledge (Name):** knowledge is extracted from a list of named entity names.
3. **Knowledge-Base-Based Knowledge (KB) :** features extracted from a KB or an entity annotated corpus.
4. **Entity-Based Knowledge (Entity):** encodes document-specific knowledge about the entities found in text to exploit the association between NER and NED

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

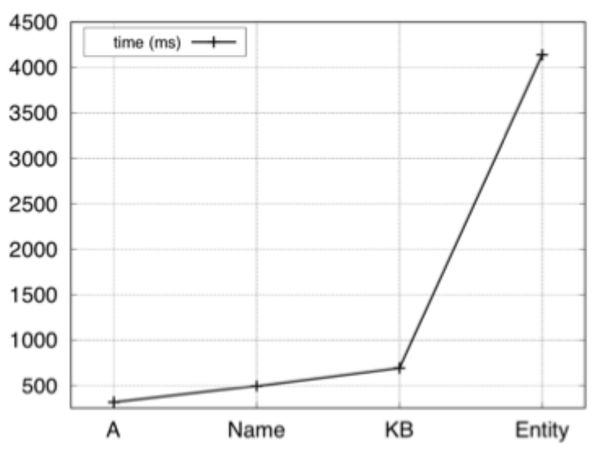
Evaluation



F1 on English Corpora



F1 on German/Spanish Corpora



Timing on CoNLL2003-test

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

Ablation Study

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.
- We showed that our method is general enough to be applied to multiple languages using automatically extracted features.

max planck institut informatik