HITS' Monolingual and Cross-lingual Entity Linking System at TAC 2013

Angela Fahrni, Benjamin Heinzerling, Thierry Göckel and **Michael Strube**
Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
{Angela.Fahrni|Benjamin.Heinzerling|Thierry.Goeckel|Michael.Strube}@h-its.org

Abstract

This paper presents HITS' system for monolingual and cross-lingual entity linking at TAC 2013. The system is an extended version of our last year's joint entity disambiguation and clustering system based on Markov Logic Networks. We describe the new extensions and discuss the results.

The results show that our approach is competitive across all three languages: with a micro-average accuracy of 0.817, our best English run is close to the one of the best system (0.833). While we had the best performing system in the Spanish cross-lingual entity linking task, our Chinese results are lower than the ones of the best system.

1 Introduction

HITS participated in the English monolingual and the Chinese and Spanish cross-lingual entity linking tasks at TAC 2013. We built upon our last year's system for joint entity disambiguation and clustering using Markov Logic Networks (Fahrni and Strube (2012) and Fahrni et al. (2013)).

Most previous work tackles the three sub-tasks involved in entity linking – entity disambiguation, recognition of NILs and clustering of NILs – using a pipeline-based approach (Ji et al., 2011). Such a pipeline-based approach leads to error propagation and fails to exploit dependencies between (intra- and cross-document) clustering and disambiguation. In Fahrni and Strube (2012), we proposed a joint system that simultaneously solves all three sub-tasks using Markov Logic Networks and showed significant improvements on ACE 2005 and TAC 2011.

Our system aims to disambiguate common and proper nouns. Instead of just disambiguating the query term, we also disambiguate other common and proper nouns that influence the decision for the query term given our features. The system is exclusively trained on the internal hyperlinks of 500 Wikipedia articles. No TAC data is used to train the system.

For the cross-lingual entity linking task, we followed our last year's strategy and used the interlanguage links to map between the languages. We only trained on English data and used the English model for the Spanish and Chinese entity linking tasks.

In the remainder of this paper, we give a brief overview of our system, explain the new features we added this year and describe our cross-lingual approach (Section 2). In Section 3, we discuss the results. Related work is presented in Section 4.

2 Approach

In this section, we give a brief overview of our joint approach for entity disambiguation and clustering and discuss the new features. For a more detailed description of the approach the reader is referred to Fahrni and Strube (2012) and Fahrni et al. (2013).

Our approach simultaneously performs disambiguation, recognition of NILs and clustering. We use Markov Logic Networks and learn all weights jointly.

Predicates	
Hidden predicates	
p1	$hasEntity(m, e)$
p2	$hasSameEntity(m, n)$
Predicate templates for disambiguation and clustering	
p3	$featureDisambiguation(m, e, s)$
p4	$featureClustering(m, n, s)$

Formulas		
Hard constraints		
f1		$\forall m \in M : \ \lvert\{e \in E : hasEntity(m, e)\}\rvert \leq 1$
f2		$\forall m, n \in M : \ m \neq n \wedge hasSameEntity(m, n) \rightarrow hasSameEntity(n, m)$
f3		$\forall m, n, l \in M : \ m \neq n \wedge m \neq l \wedge n \neq l$ $\wedge\ hasSameEntity(m, n) \wedge hasSameEntity(n, l) \rightarrow hasSameEntity(m, l)$
f4		$\forall m, n \in M : \ m \neq n \wedge hasSameEntity(m, n) \wedge hasEntity(m, e)$ $\rightarrow hasEntity(n, e)$
f5		$\forall m, n \in M : \ m \neq n \wedge hasEntity(m, e) \wedge hasEntity(n, e)$ $\rightarrow hasSameEntity(m, n)$
Template formulas with learned weights		
f6	$(w \cdot s)$	$\forall m \in M\ \forall e \in E_m : \ featureDisambiguation(m, e, s)$ $\rightarrow hasEntity(m, e)$
f7	$(w \cdot s)$	$\forall m, n \in M : \ m \neq n \wedge featureClustering(m, n, s)$ $\rightarrow hasSameEntity(m, n)$

Table 1: Predicates and formulas used for entity disambiguation and clustering (m, n, l represent mentions, M sets of mentions, e an entity, E all entities, E_m all candidate entities for mention m and s scores)

2.1 Markov Logic Networks

Markov Logic (ML) combines first-order logic with probabilities (Domingos and Lowd, 2009). A Markov Logic Network (MLN) is a first-order knowledge base and consists of a set of pairs (F_i, w_i), where F_i is a first-order formula and $w_i \in \mathbb{R}$ is the weight of formula F_i. It is a template for constructing a Markov Network. This Markov Network has a binary node for each possible grounding for each predicate of the MLN. If the grounding of the predicate is true, the binary node's value is set to 1, otherwise to 0. Furthermore, it contains one feature[1] for each ground formula F_i. If a ground formula is true, its feature's value is set to 1, otherwise to 0. The feature's weight is provided by w_i.

The probability distribution in the ground Markov Network is given by

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(x)\right)$$

where $n_i(x)$ is the number of true groundings of F_i in x. The normalization factor Z is the partition function.

To perform MAP inference we use *thebeast*[2] which transforms the inference problem into an Integer Linear Program and solves it using cutting plane inference (Riedel, 2008). To learn the weights we use a Perceptron.

2.2 Disambiguation and Clustering with MLNs

Entity disambiguation and clustering are two different ways to deal with lexical ambiguities. The two tasks focus on different relations:

[1] In this section *feature* is used differently than in the rest of the paper.

[2] `http://code.google.com/p/thebeast`.

ID	Predicates	Description
		Local Features
i1	$hasPriorProbability(m, e, s)$	The prior probability is defined as the probability that a mention m refers to an entity e and is estimated based on the English, Chinese and Spanish Wikipedia dump respectively.
i2	$hasRelatedness(m, e, s)$	This feature reflects the average pairwise relatedness of a candidate entity for a mention to the context. The whole text serves as context.
i3	$hasCoocProbability(m, e, s)$	This feature measures the average co-occurrence probability of a candidate entity given a mention and its context.
i4	$hasContextSimilarity(m, e, s)$	The local context similarity measures how similar the current local context C_m is to the local contexts for that entity in Wikipedia.
i5	$hasStringDistance(m, e, s)$	This feature accounts for the difference between the mention string m used in the text and the preferred name p for a candidate entity of m.
i6	$(\star)hasDescriptorNeighbours(m, e, s)$ $(\star)hasDescriptorSentence(m, e, q)$ $(\star)hasDescriptorDocument(m, e, q)$	Article titles in Wikipedia often contain domain descriptors in brackets or after a comma. For each candidate entity of a mention, all preferred names are obtained (redirects and article titles) and all domain descriptors are extracted from these names. It is then checked if some of these descriptors occur in the intermediate context of a mention (context window of 4), in the same sentence or in the same document. The score s is the relative proportion of domain descriptors for a candidate entity given the domain descriptors for all candidate entities for a mention.
		Intra-document Clustering Features
i7	$haveSameLemma(m, n, s)$	The one sense per discourse assumption states that one mention string is used to refer to one sense, i.e. in our case to one entity, in one document (Gale et al., 1992).
i8	$isSubStringHeadMatch(m, n, s)$	The one entity per discourse assumption often applies to mentions which are sub-strings of each other and share the same syntactic head lemma.
i9	$isPartialStringMatch(m, n, s)$	If two mentions are person names and one is a sub-string of the other, we assume that they refer to the same entity with a certain probability.
i10	$(\star)isCoreferent(m, n, s)$	If two mentions are coreferent according to Stanford's coreference resolution system, we assume that they refer to the same entity with a certain probability. The score s is the inverse distance between the two mentions in sentences.
i11	$(\star)isAcronym(m, n, s)$	In newspaper articles, acronyms are quite common. In a first step all acronyms are identified. A mention is considered as an acronym if it contains at most seven characters, starts with an uppercase letter and also contains uppercase letters within the word. In the second step, all multi-word mentions are identified and decomposed into words. For each token of a multi-word mention, we take the first letter and form a pseudo-acronym. If a pseudo-acronym matches an acronym identified in the previous step, an isAcronym relation is established between the two mentions. The score s is the inverse distance in sentences. If for an acronym no corresponding pseudo acronym is identified, longer mentions in the same document with at least one common candidate entity are considered as the full version of an acronym. In our Wikipedia training data, acronyms are relatively rare. Hence it is difficult to learn a weight for the acronym feature. As it is similar to the *isSubStringHeadMatch* feature, we use the same weight for the two features.

Table 2: Features for disambiguation and intra-document clustering. m, n denote mentions, e an entity, s a score. The predicates are plugged in the template formulas $f6$ and $f7$ in Table 1.

ID	Predicates	Description
		Cross-document Clustering Features
i12	$(\star)hasMentionOverlap(m, n, s)$	Each text is represented as a vector containing all named entities that appear in this text. The value of all vector entries is set to 1. If two mentions m and n belong to two different documents and have the same string, we calculate the cosine similarity between the respective document vectors. If this cosine similarity is higher than a certain threshold (the threshold is set to 0.2), we establish an *hasMentionOverlap*-relation with the cosine similarity as a score. The assumption is that if two documents are similar with respect to named entities, the two mentions with the same string are likely to refer to the same entity.
i13	$(\star)hasEntityOverlap(m, n, s)$	For each text, an entity-based representation is obtained by taking all candidate entities with a prior probability of at least 0.95. These selected entities form a vector representation of a text. The value of each entry is set to 1. If two mentions m and n belong to two different documents and have a similar string (the string edit distance has to be smaller than 0.2), we calculate the cosine similarity between the respective document vectors. If this cosine similarity is higher than a certain threshold (the threshold is set to 0.1), we establish an *hasEntityOverlap*-relation with the cosine similarity as a score. In contrast to the *hasMentionOverlap*-feature, the document representation is language-independent and can also be used for cross-lingual clustering. In addition, this representation abstracts away from the surface forms and can deal with synonymy.
i14	$(\star)hasCategoryOverlap(m, n, s)$	Given the entity-based representation (cf. *hasEntityOverlap*), a document representation is established by taking the categories associated with the Wikipedia articles of the selected entities. We also exploit the category hierarchy and consider a category if it is reachable within three steps given an article. If two mentions m and n belong to two different documents and have a similar string (the string edit distance has to be smaller than 0.2), we calculate the cosine similarity between the respective document vectors. If this cosine similarity is higher than a certain threshold (the threshold is set to 0.3), we establish an *hasCategoryOverlap*-relation with the cosine similarity as a score. The document representation is language-independent and fits the cross-lingual clustering requirements. The category-based representation is more noisy than the entity-based one, but may allow to capture domain similarities that can not be captured by the entity-based representation.
i15	$(\star)stringCompatibility(m, n)$	This feature is a more restrictive cross-document string match feature that takes into account modifier mismatches between longer mentions that are in an intra-document clustering relation with the target mentions. Given two mentions m and n from two different documents, we consider them as similar if either one mention is a sub-string of the other mention or if the length-normalized edit distance between the two mentions is at most 0.2. If the two mentions are similar, we obtain all mentions that are in an intra-document clustering relationship with the respective mentions given our intra-document clustering features. For each document we select the longest mention that is in an intra-document clustering relationship (given by our intra-document clustering features) with the target mention. We establish a *stringCompatibility*-relationship between m and n if the respective selected longest mentions share some candidate entities, have a string similarity higher than 0.9, are sub-strings of each other or are acronyms.

Table 3: Features for cross-document clustering. m, n denote mentions, s a score. The predicates are plugged in the template formula $f7$ in Table 1.

- Entity disambiguation models the relation between mentions and entities in a knowledge base. In order to solve lexical ambiguities, mentions are linked to entities in a given knowledge base. If the referred entity is not part of the knowledge base, we consider the mention as a NIL.

- Entity clustering models the relation between mentions. Mentions are clustered, so that all mentions in a cluster refer to the same entity.

For both relations to be predicted a hidden predicate is defined: *hasEntity(Mention, Entity)* models a relation between a mention and an entity (disambiguation); *hasSameEntity(Mention, Mention)* models the relation between two mentions that refer to the same entity (clustering). The relations between the two tasks are defined with hard constraints. The hidden predicate and the hard constraints build the core of our approach and are given in Table 1. M denotes all mentions and E_m refers to all candidate entities of a mention m. Table 1 also contains two template formulas with learned weights (Template Formulas $f6$ and $f7$). These two template formulas serve as a building block for features (see Section 2.3) and need to be instantiated. The final weight of the respective instantiated formulas is given by the learned weight w and a score s (e.g. the prior probability).

2.3 Features

Table 3 summarizes all features. The features marked by a star ($\star$) changed compared to last year or are newly introduced this year. All other features are described in more details in Fahrni and Strube (2012) and Fahrni et al. (2013).

All features have a corresponding predicate which is part of at least one formula (see Template Predicate $p3$ and $p4$ and Template Formulas $f6$ and $f7$ in Table 1).

2.4 Cross-lingual Entity Linking

We followed our last year's strategy and try to do as little language adaptation as possible in order to investigate how portable our current system is.

We map the Chinese and Spanish Wikipedia articles to the English articles using inter-language links[3] and disambiguate Chinese and Spanish mentions directly with respect to these mapped articles. In case a Chinese or Spanish article can not be mapped, we assign it a new ID.

The only parts that are language-specific are the preprocessing and the lexicon, which we extracted from the Wikipedia dump in the respective language.

Both the monolingual and the cross-lingual system are trained on 500 English Wikipedia articles.

3 Experiments

3.1 Processing TAC queries

Our preprocessing pipeline consists of the following four steps:

1. **Text cleaning:** HTML tags and noise (e.g. in Web documents) are removed.
2. **Preprocessing:** We tokenize the texts and perform POS tagging and parsing. For English and Chinese, we use *Stanford's CoreNLP*[4], for Spanish, we use *FreeLing* (Padró and Stanilovsky, 2012).
3. **Mention detection and feature extraction:** In this step mentions are identified and the features extracted. To identify candidate entities for the mentions we use a lexicon. To create this lexicon we extract all anchors, article titles and titles from redirects from the English, Chinese and Spanish Wikipedia dumps. In order to reduce the noise of anchors in English, we only consider an anchor if it is used at least two times to refer to the respective Wikipedia article. We did not set this constraint in Chinese and Spanish, as these Wikipedia dumps are smaller.
4. **Inference:** We use *thebeast* (Riedel, 2008). We disambiguate and cluster the query terms as well as all identified mentions that influence the decision for the query terms given our features directly or transitively. As the clustering can cross document boundaries, we process all mentions from all documents that can be in the same cluster (given our features) at the same time.

[3] We use the following Wikipedia dumps: English (2012/01/04), Chinese (2012/08/22), Spanish (2012/07/28), German (2012/01/16), Italian (2012/01/26), Dutch (2012/01/19).

[4] http://nlp.stanford.edu/software/corenlp.shtml

Run ID	Features / Formulas	Postprocessing	Resources
English Entity Linking Task			
ENHITS1	full model (except $i13$, $i14$, $i15$)	none	*CoNLL* gender data for partial string match feature
ENHITS2	full model (except $i13$, $i14$, $i15$)	unclustered NILs: clustering based on distributional features	*CoNLL* gender data for partial string match feature
ENHITS3	full model (except $i13$, $i14$, $i15$)	unclustered NILs: clustering based on distributional features and string match	*CoNLL* gender data for partial string match feature
ENHITS4	full model (except $i10$, $i13$, $i14$, $i15$)	none	*CoNLL* gender data for partial string match feature
ENHITS5	full model (except $i10$, $i13$, $i14$, $i15$)	clustering of joint approach ignored, all NILs as singletons	*CoNLL* gender data for partial string match feature
Chinese Entity Linking Task			
ZHHITS1	full model, without $i10$ and $i12$	none	List from *Baidu Baike* for partial string match feature
ZHHITS2	full model, without $i10$, $i12$ and $i15$	none	List from *Baidu Baike* for partial string match feature
ZHHITS3	full model, without $i10$, $i12$, $i14$ and $i15$	none	List from *Baidu Baike* for partial string match feature
ZHHITS4	full model, without $i10$, $i13$, $i14$ and $i15$	none	List from *Baidu Baike* for partial string match feature
ZHHITS5	full model, without $i10$ and $i13$, $i14$, $i15$	unclustered NILs: string match clustering	List from *Baidu Baike* for partial string match feature
Spanish Entity Linking Task			
ESHITS1	full model, without $i10$ and $i12$	none	
ESHITS2	full model, without $i10$, $i12$ and $i15$	none	
ESHITS3	full model, without $i10$, $i12$, $i14$ and $i15$	none	
ESHITS4	full model, without $i10$, $i13$, $i14$ and $i15$	none	
ESHITS5	full model, without $i10$, $i13$, $i14$ and $i15$	unclustered NILs: string match clustering	

Table 4: Description of the different runs of HITS for the monolingual and cross-lingual entity linking tasks at TAC 2013.

Run	Micr.	B^3 P	B^3 R	B^3 F1	B^{3+} P	B^{3+} R	B^{3+} F1
English Entity Linking Task							
Best	0.833						0.746
Median	0.746						0.574
ENHITS1	**0.817**	**0.954**	**0.662**	**0.782**	**0.792**	**0.601**	**0.684**
ENHITS2	**0.817**	**0.954**	**0.662**	**0.782**	**0.792**	**0.601**	**0.684**
ENHITS3	0.817	0.923	0.665	0.773	0.766	0.603	0.675
ENHITS4	**0.817**	**0.954**	**0.662**	**0.782**	**0.792**	**0.601**	**0.684**
ENHITS5	0.817	0.966	0.569	0.716	0.802	0.509	0.623
Chinese Cross-lingual Entity Task							
Best	0.815						0.667
ZHHITS1	0.779	0.750	0.689	0.718	0.611	0.589	0.600
ZHHITS2	0.781	0.827	0.637	0.720	0.654	0.555	0.600
ZHHITS3	**0.780**	**0.834**	**0.648**	**0.729**	**0.656**	**0.566**	**0.607**
ZHHITS4	0.777	0.933	0.536	0.681	0.721	0.485	0.580
ZHHITS5	0.777	0.781	0.643	0.706	0.625	0.557	0.589
Spanish Cross-lingual Entity Task							
Best	0.815						0.709
ESHITS1	0.774	0.814	0.829	0.821	0.697	0.658	0.677
ESHITS2	**0.815**	**0.930**	**0.760**	**0.837**	**0.781**	**0.649**	**0.709**
ESHITS3	0.812	0.927	0.756	0.833	0.775	0.644	0.704
ESHITS4	0.798	0.942	0.726	0.820	0.772	0.612	0.683
ESHITS5	0.798	0.910	0.736	0.814	0.748	0.619	0.678

Table 5: HITS' performance compared to the best and median scores in the monolingual and cross-lingual entity linking tasks.

3.2 Settings

HITS participated in all three entity linking tasks. In total, we submitted five runs for each sub-task. The runs differ in features used and whether or not post-processing was applied. Table 4 describes the differences between the runs. The 'post-processing' column indicates if we submitted the output of our joint entity disambiguation and clustering system without any modifications (none), or if we performed additional NIL-clustering in a post-processing step. *String match clustering of NILs* means that all NILs that are singletons after the inference step are clustered using a string match heuristic (ZHHITS5, ESHITS5). For English, we experimented with distributional features in a post-processing step. Since the local contexts of NIL mentions were not informative because of data sparseness, we clustered NILs based on the following similarity measures between the documents containing the NILs:

- Cosine similarity based on a 1000-dimensional vector space model.[5]
- Similarity of topic distributions: After creating a topic model with 50 topics using Mallet [6], we computed the Hellinger Distance between the topic distributions of each document pair.

However, this additional NIL clustering step did not improve performance (ENHITS2, ENHITS3).

3.3 Results

Table 5 shows the results for all runs. The best runs are highlighted in bold and are produced without any postprocessing. The results indicate that our approach is competitive. For English, the accuracy (micro-average score) is very close to the one of the best system and well above the medium result. The

[5]Using standard settings of the SemanticVectors package (http://code.google.com/p/semanticvectors/)

[6]http://mallet.cs.umass.edu/

Run	ENG		ES/ZH		Overall	
	Micr.	**B^3 F1**	**Micr.**	**B^3 F1**	**Micr.**	**B^3 F1**
GPE						
Best EN	0.829	0.746			0.829	0.746
ENHITS1	0.822	0.734			0.822	0.734
Best ZH	*0.972	*0.896	0.858	0.762	*0.878	0.790
ZHHITS3	0.938	0.794	0.855	0.736	0.875	0.750
Best ES	*0.915	*0.865	*0.771	*0.728	*0.815	*0.772
ESHITS2	0.915	0.862	0.757	0.716	0.808	0.762
PER						
Best EN	0.847	0.778			0.847	0.778
ENHITS1	0.816	0.684			0.816	0.684
Best ZH	*0.914	0.729	0.671	0.559	0.727	0.601
ZHHITS3	0.897	0.575	0.521	0.379	0.613	0.434
Best ES	0.912	0.731	0.819	0.635	*0.830	0.650
ESHITS2	0.871	0.680	0.817	0.629	0.830	0.642
ORG						
Best EN	0.884	0.737			0.884	0.737
ENHITS1	0.812	0.618			0.812	0.618
Best ZH	0.854	0.767	0.877	0.726	0.871	0.732
ZHHITS1	0.823	0.739	0.853	0.580	0.846	0.620
Best ES	*0.834	*0.762	*0.801	*0.703	*0.808	*0.718
ESHITS2	0.829	0.755	0.801	0.703	0.808	0.718

Table 6: Results per named entity type. We report the best reported results (*Best EN, Best ZH, Best ES*) and the results of our best runs (*ENHTS1, ZHHITS1, ESHITS2*). Note that the best reported numbers for a language are not necessarily from the same system, e.g. the system with the best micro-average accuracy is not necessary the one with the best B^3F1 score. The star (*) in front of some results in *Best* indicate that it is achieved by one of our runs. For each entity type, we report the results for the English queries and the Chinese/Spanish queries separately as well as the overall scores.

results for Chinese are lower than the ones of the best system. In the Spanish cross-lingual sub-task our approach outperforms the other systems by more than 5% in terms of B^3F1.

A more detailed analysis of the English results reveals that adding coreference information does not improve the results (ENHITS1 vs. ENHITS4). One reason for the lack of any improvement is that the most important features for noun-noun coreference resolution are already integrated in the system (string match, head match, acronym resolution). Hence, integrating coreference relations does not add any new information.

For all languages, the precision (B^3P) is much higher than the recall (B^3R). This behavior comes from the fact that our clustering is rather conservative and only makes use of high precision features.

Table 6 shows for each sub-task the results per entity type (GPE, PER, ORG) and language (ENG, ZH, ES). While some of our runs outperform our overall best run for a specific entity type, we only report the results of our overall best run for each sub-task. As Table 6 shows, our results for GPEs are highly competitive across all languages in terms of both micro-average accuracy and B^3F1 score. For PERs and ORGs the accuracy is still high, while the B^3F1 score is lower. This is mainly due to relatively low recall scores for clustering. The lower results in the Chinese sub-task are due to low results for Chinese person names. Inspection of some examples reveals that our system currently fails to capture small variations in Chinese names, e.g. due to variations in transliterations of foreign names such as the ones for *Susan Rice*. A few examples are shown in Fig-

ZH Translit.	ZH Translit.	English
苏珊·赖 斯	苏珊・莱斯	Susan Rice
史蒂夫・欧文	史蒂夫·艾尔文	Steve Irwin
威廉姆斯	威廉斯	Williams

Figure 1: Examples of variations in Chinese transliterations of English names.

ure 1. This also affects the cross-lingual clustering. While one of the Chinese variations for *Susan Rice* is clustered with the corresponding English query, the other one is part of a different cluster.

4 Related Work

Most work on entity linking approaches the task of entity disambiguation, recognition of NILs and clustering using a pipeline-based approach (Ji et al., 2011). Monahan et al. (2011) interleave entity linking and clustering, but they do not approach the two tasks jointly: after disambiguation, mentions are clustered. Then each cluster is assigned an entity in the knowledge base if there exists a corresponding one. Dai et al. (2011) perform entity disambiguation and recognition of the NILs jointly using Markov Logic Networks. In contrast to us, they do not cluster mentions and focus on one specific type of mentions in the biological domain, namely mentions that refer to genes. To our knowledge, our last year's system (Fahrni and Strube, 2012; Fahrni et al., 2013) is the first joint approach for all three sub-tasks.

While early work often uses local classifiers or rankers that select an entity for each mention independently (Csomai and Mihalcea, 2008; Milne and Witten, 2008; Dredze et al., 2010), recently various global approaches have been proposed. Kulkarni et al. (2009) propose a method that maximizes local context-concept compatibility and global concept coherence. Han and Sun (2012) use a generative model integrating topic coherence (one topic per document) and local context compatibility. Ratinov et al. (2011) describe a two pass method and use the output of the first pass as input for the second one. While all these approaches use a limited number of global features, we integrate various global features and also learn their weights. The closest approach to ours is the one of Cheng and Roth (2013). They also integrate intra-document coreference information, but do not make use of cross-document clustering features.

5 Conclusions

HITS participated with an extended version of last year's approach in the monolingual English and the Chinese and Spanish cross-lingual entity linking tasks. Our approach performs joint disambiguation and clustering using Markov Logic Networks. The results are competitive across languages.

Acknowledgments. This work has been partially funded by the Klaus Tschira Foundation.

References

Xiao Cheng and Dan Roth. 2013. Relational inference for Wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Wash., 18–21 October 2013, pages 1787–1796.

Andras Csomai and Rada Mihalcea. 2008. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems*, 23(5):34–41.

Hong-Jie Dai, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2011. Entity disambiguation using a Markov-Logic network. In *Proceedings of the 5th International Joint Conference on Natural Language Processing,* Chiang Mai, Thailand, 8–13 November 2011, pages 846–855.

Pedro Domingos and Daniel Lowd. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan Claypool Publishers.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pages 277–285.

Angela Fahrni and Michael Strube. 2012. Jointly disambiguating and clustering concepts and entities with Markov logic. In *Proceedings of the 24th International Conference on Computational Linguistics,* Mumbai, India, 8–15 December 2012, pages 815–832.

Angela Fahrni, Thierry Göckel, and Michael Strube. 2013. HITS' monolingual and cross-lingual entity linking system at TAC 2012: A joint approach. In *Proceedings of the Text Analysis Conference,* National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language*

Workshop, New York, N.Y., 23-26 February 1992, pages 233–237.

Xianpei Han and Le Sun. 2012. An entity-topic model for entity linking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics,* Jeju Island, Korea, 8–14 July 2012, pages 105–115.

Heng Ji, Ralph Grishman, and Hoa Dang. 2011. Overview of the TAC 2011 knowledge base population track. In *Proceedings of the Text Analysis Conference,* National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* Paris, France, 28 June – 1 July 2009, pages 457–466.

David Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM 2008),* Napa Valley, Cal., USA, 26–30 October 2008, pages 1046–1055.

Sean Monahan, John Lehmann, Timothy Nyberg, Jesse Plymale, and Arnold Jung. 2011. Cross-lingual cross-document coreference with entity linking. In *Proceedings of the Text Analysis Conference,* National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: towards a wider multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation,* Istanbul, Turkey, 21–27 May 2012.

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,* Portland, Oreg., 19–24 June 2011, pages 1375–1384.

Sebastian Riedel. 2008. Improving the accuracy and efficiency of MAP inference for Markov logic. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence,* Helsinki, Finland, 9–12 July 2008, pages 468–475.