A Latent Variable Model for Discourse-aware Concept and Entity Disambiguation

Angela Fahrni and Michael Strube
Heidelberg Institute for Theoretical Studies gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
(angela.fahrni|michael.strube)@h-its.org

Abstract

This paper takes a discourse-oriented perspective for disambiguating common and proper noun mentions with respect to Wikipedia. Our novel approach models the relationship between disambiguation and aspects of cohesion using Markov Logic Networks with latent variables. Considering cohesive aspects consistently improves the disambiguation results on various commonly used data sets.

1 Introduction

> *"I have to review a **paper**", the **supervisor** moaned from the **office**. "Please don't disturb me until I'm done with the **review**." His **student** nodded, went to the **cafeteria**, sat down in the **sunshine** and started to read yesterday's **paper**.*

This text snippet illustrates two aspects that have been neglected by previous disambiguation approaches. (1) The interpretation of different mentions, i.e. common and proper nouns, is determined by different notions of context: some mentions depend more on a local sentence-level context (*paper* in *read yesterday's paper*; the global context is misleading), some more on a global one (*review* in *I'm done with the review*; the local context is not discriminative), some on both global and local context (*paper* in *review a paper*). (2) The context relevant to disambiguate a mention depends on how it is embedded into discourse and is not bound to the surface form of a mention (*paper* in the first sentence vs. *paper* in the last one).

Starting from this observation, we argue that the context relevant to disambiguate a mention correlates with its cohesive scope, i.e. the text span within which a mention establishes cohesive relations. Therefore, we propose to disambiguate mentions differently depending on their cohesive scopes (Section 2). We distinguish between three different cohesive scopes of mentions and model them as latent variables using Markov Logic Networks (Section 3). The use of latent variables allows us to learn and predict the cohesive scope and the disambiguation of a mention jointly. This comes with the advantage that the learning of the scope assignment does not need annotated data by itself but is guided by the annotations available for the target prediction task, i.e. the disambiguation.

In this paper, we focus on concept and entity disambiguation[1] with respect to an inventory derived from Wikipedia and compare (1) to a state-of-the-art approach that treats all mentions alike and uses the same features for disambiguation, (2) to a pipeline-based approach, and (3) to other state-of-the-art approaches (Section 4).

While early work disambiguated concepts using the local context (Csomai and Mihalcea, 2008), current research focuses on exploiting the global document context (Milne and Witten, 2008; Kulkarni et al., 2009; Ratinov et al., 2011; Fahrni and Strube, 2012; Cheng and Roth, 2013). Although such global approaches try to balance between local and global context, they treat all mentions alike, i.e., they apply the same model and the same weighting of local and global context features for disambiguating all mentions (Section 5).

2 Motivation

Halliday and Hasan (1976) define *cohesion* as "relations of meaning that exist within the text, and that define it as a text" (p. 4). A *tie* is one instance of such a cohesive relation between two items. Cohesive ties occur on various linguistic levels, such as on the entity level (e.g. coreference and bridging relations) or on the concept level (e.g. lexical

[1] In the following, we use *concept* to refer to concepts and what is usually called entities (e.g. Ji et al. (2011)).

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 491–500,
Gothenburg, Sweden, April 26-30 2014.

chains). In this paper, we focus on concept-level cohesion and assume that each concept referred to by a mention can exhibit cohesive ties with concepts from other lexical units. The *cohesive scope* of a mention is the text span within which a concept referred to by a mention shows such cohesive ties. We distinguish three broad categories of cohesive scopes: (1) Mentions with *local cohesive scope* exhibit cohesive ties with lexical units in the same sentence; (2) mentions with *intermediate cohesive scope* show cohesive ties both within the sentence and beyond; (3) mentions with *global cohesive scope* form cohesive ties with mentions across sentence boundaries.

The notion of scope is a means to define the appropriate context to disambiguate a mention. A mention of local scope does not exhibit relations with lexical units outside its sentence. Hence, the global context does not help to disambiguate it or can even lead to the wrong disambiguation. For a mention with global scope, the global context is crucial, while the local context is not discriminative or even misleading. For a mention with intermediate scope both local and global context are relevant. Hence, while the scope influences the appropriate disambiguation context, the disambiguation of a mention influences its scope. In the example (Section 1), *paper* in *read yesterday's paper* refers to the concept NEWSPAPER. Its scope is local, as it lacks some cohesive ties with mentions in other sentences. If it had been disambiguated to SCHOLARLY PAPER, its scope would be global. This reciprocal relationship between discourse structure and meaning has also been discussed by Asher and Lascarides (1995). They use rhetorical relations for structuring discourse while we rely on the notion of lexical cohesion and model scope assignment and disambiguation jointly.

Our notion of scope is related to work on lexical chains (Morris and Hirst, 1991; Nelken and Shieber, 2006; Mihalcea, 2006) and to work in content modeling, e.g. Haghighi and Vanderwende (2009) distinguish content vocabulary and document-specific vocabulary.

3 Approach

Given a set of features for disambiguation, we aim to weight them differently depending on the scope. To model the reciprocal relationship between scope assignment and disambiguation, we propose a latent variables based approach using Markov Logic Networks that allows us to learn the parameters for the scope assignment and the disambiguation tasks jointly and enables us to perform joint inference.

Our approach is *joint* as we assign the scope s and predict the concept c for a mention m simultaneously. As during learning training data is available for the disambiguation task but not for the scope assignment task, we face a problem with *latent variables*. Latent variables represent missing information in the input or a part of the output which is not relevant except for supporting the prediction of the target (Smith, 2011). In our approach, the different cohesive scopes are modeled by latent variables. Each mention to be disambiguated is assigned a scope s. All feature weights are parametrized by scope s. The parameters for the disambiguation and scope assignment tasks are learned jointly and are guided by the annotations available for the disambiguation task.

Markov Logic Networks can be represented as log-linear models, when grounded, and are therefore straightforward to extend with latent variables (Smith, 2011; Poon and Domingos, 2008). In addition, global features can be conveniently integrated.

3.1 Markov Logic Networks

Markov Logic (ML) incorporates first-order logic and 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[2] 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

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

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

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*[3] which transforms the inference problem into an Integer Linear Program and solves it using cutting plane inference (Riedel, 2008).

3.1.1 Weight Learning with Latent Variables

Since no annotations are available for the scope distinction, we face a latent variable learning problem. For learning weights in this situation we follow Poon and Domingos (2008). We split our hidden predicates into two parts: V are the ones for which the ground truth is known (concepts) and U are the ones for which there is no annotation (scopes). Let O be the observed predicates. Let o and v be the values of O and V in the training data. u denotes values assigned to U. Weight learning finds a w that maximizes the conditional log-likelihood

$$\begin{aligned} L_w(o,v) &= \log P_w(V=v|O=o) \\ &= \log \sum_u P_w(V=v,U=u|O=o), \end{aligned}$$

where the sum is over all possible values of U.

Although $L_w(o,v)$ is not convex, a local optimum can be found via gradient descent by iteratively solving

$$w_{t+1} = w_t + \eta \nabla_w L_w(o,v),$$

where the gradient $\nabla_w L_w(o,v)$ is given by

$$\frac{\partial}{\partial w_i} L_w(o,v) = E_w[n_i(o,v,U)] - E_w[n_i(o,V,U)].$$

E_w denotes the expectation according to P_w and $n_i(o,v,u)$ is the number of true groundings of formula F_i under the assignment specified by (o,v,u). We use a voted perceptron (Lowd and Domingos, 2007) which approximates the expectations via computing the MAP solution with (o,v) fixed ($E_w[n_i(o,v,U)]$) and (o) fixed ($E_w[n_i(o,V,U)]$) respectively.

3.1.2 Scope-aware Concept Disambiguation

Both the scope assignment and the disambiguation task are performed jointly using Markov Logic Networks.

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

Table 1 shows the core of our proposed approach in terms of predicates and first-order logic formulas. We build upon our previous approach for joint concept disambiguation and clustering (Fahrni and Strube, 2012). For brevity, we only discuss the scope-aware extension of the disambiguation part. The extension for clustering is done analogously.

The purpose of assigning a scope to each mention m is to learn scope-specific weights for disambiguation to account for heterogenous scopes of mentions. The learned weights are parametrized by scopes. We indicate this parametrization of learned weights by $w(s)$ (cf. Table 1, $f8$, $f9$).

For each relation to predict, a hidden predicate is defined. We are interested in predicting two relations: a relation between a mention m and a concept c ($p1$: *hasConcept(m, c)*) and a relation between a mention m and a scope s ($p3$: *hasScope(m, s)*). To bridge between the disambiguation and the scope assignment task a third hidden predicate *relatesScopeToConcept(m, c, s)* ($p2$) models a relation between a mention m, a concept c and a scope s. This predicate together with Formulas $f4 - f7$ garantuees that the scope assignment and the selection of a concept for a mention influence each other and that the ground hidden predicates are in accordance.[4] Hard cardinality constraints (*f1*, *f2*, *f3*) enforce that each mention m is assigned exactly one scope s and at most one concept c.

The hidden predicates and formulas form the core. Features for the disambiguation and the scope assignment tasks are incorporated using local and global formulas with learned weights. The features are described in Section 3.2. Table 1 gives formula templates for both tasks (please note that these are templates not formulas (Section 3.2)): (1) a template for formulas that add information for scope assignment ($f8$) and (2) a template for formulas that add information for disambiguation ($f9$). All formulas with scope-parametrized weights that are relevant for the concept prediction task are defined for the predicate *relatesScopeToConcept*. This enables us to activate the relevant

[4] We also run experiments with just two hidden predicates, i.e. *hasConcept(m, c)* and *hasScope(m, s)*. All formulas with learned weight were then defined in the following, less efficient way: $\forall m \in M, c \in C, s \in S$: $featureDisambiguation(m,c,q) \rightarrow hasConcept(m,c) \wedge hasScope(m,s)$. q is a score (Table 1).

Predicates	
Hidden predicates	
p1	$hasConcept(m, c)$
p2	$relatesScopeToConcept(m, c, s)$
p3	$hasScope(m, s)$
Predicate template for disambiguation features	
p4	$featureDisambiguation(m, c, q)$
Predicate template for scope assignment features	
p5	$featureScope(m, q)$
Formulas	
Hard cardinality constraints	
f1	$\forall m \in M: \ \lvert\{c \in C : hasConcept(m, c)\}\rvert \leq 1$
f2	$\forall m \in M: \ \lvert\{c \in C, s \in S : relatesScopeToConcept(m, c, s)\}\rvert \leq 1$
f3	$\forall m \in M: \ \lvert\{s \in S : hasScope(m, s)\}\rvert = 1$
Hard constraints	
f4	$\forall m \in M, c \in C, s \in S: \ relatesScopeToConcept(m, c, s) \rightarrow hasConcept(m, c)$
f5	$\forall m \in M, c \in C, s \in S: \ relatesScopeToConcept(m, c, s) \rightarrow hasScope(m, s)$
f6	$\forall m \in M, c \in C, s \in S: \ hasConcept(m, c) \wedge hasScope(m, s) \rightarrow relatesScopeToConcept(m, c, s)$
f7	$\forall m \in M, c \in C: \ hasConcept(m, c) \rightarrow (\lvert\{s \in S : relatesScopeToConcept(m, c, s)\}\rvert = 1)$
Formula template with learned weights for scope assignment	
f8	$q \cdot w(s) \quad \forall m \in M, s \in S: \ featureScope(m, q) \rightarrow hasScope(m, s)$
Formula template with learned weights for disambiguation	
f9	$q \cdot w(s) \quad \forall m \in M, c \in C, s \in S: \ featureDisambiguation(m, c, q) \rightarrow relatesScopeToConcept(m, c, s)$

Table 1: Predicates and formulas used for scope distinction and disambiguation (m represents a mention, M sets of mentions, c a concept, C sets of concepts, s a scope, S sets of scopes, q scores, w weights and $w(s)$ a weight which is parametrized by s). The two template predicates and formulas are generalized patterns to integrate the features for the scope assignment and disambiguation task (Section 3.2).

scope-specific weights $w(s)$ which depend on the chosen scope s. The final weight for a formula can also include a score q defined by the observed predicate.

3.2 Features

For disambiguation and clustering we build upon our previous work (Fahrni and Strube, 2012). We use the same features and formulas and adopt the latter to learn scope-specific weights. Given for example the local context similarity feature (predicate *hasContextSimilarity(m, c, q)* where q is the similarity score) and the corresponding formula

$$\forall m \in M, c \in C_m: \ hasContextSimilarity(m, c, q) \rightarrow hasConcept(m, c)$$

with weight $(q \cdot w)$ we adopt it in the following way (cf. Table 1, template $f9$):

$$\forall m \in M, s \in S, c \in C_m: \ hasContextSimilarity(m, c, q) \rightarrow relatesScopeToConcept(m, c, s)$$

with weight $(q \cdot w(s))$.

In order to distinguish between the three proposed scopes, we use the features described in Table 2. The first column shows the predicate which can be used for template $f8$ in Table 1.

4 Experiments

We compare our novel scope-aware approach to our previous scope-ignorant approach (Fahrni and Strube, 2012) – which has achieved good results in the English monolingual and Chinese and Spanish cross-lingual entity linking tasks at TAC 2012 and 2013 (Fahrni et al., 2014) – and a scope-aware pipeline-based approach using the same features and preprocessing to ensure a fair comparison. This allows us to identify the differences in the results that are due to scope-awareness and differences in the results that are due to different learning strategies (joint vs. pipeline-based). In addition, we compare our joint scope-aware approach to state-of-the-art approaches using various data sets.

4.1 Data

Table 3 summarizes our test sets (ACE 2005, ACE 2004, MSNBC and TAC 2011) and our training and development sets derived from Wikipedia (WP Training, WP Dev). For each data set we report the total number of annotated mentions, the number of mentions with a corresponding concept in Wikipedia (non-NILs) and the number of NILs (i.e. mentions that do not refer to a Wikipedia con-

Predicates	Description
Mention-based Features	
$idfHead(m, q)$	The more frequent a mention is, the more likely it is to exert a local scope. This is inspired by work on indexing for IR. We use the idf score of the head of a mention according the English Gigaword Corpus (Parker et al., 2011).
$propernoun(m)$	Proper nouns are usually more prominent than common nouns and are more likely to have an intermediate or global scope than common nouns.
$singlewordNoun(m)$	Single word NPs are often less prominent than multi-word NPs and are more likely to be of local scope.
$abbrev(m)$	Abbreviations with a terminal dot such as *Mr.* or *Ltd.* tend to have a local scope as they are usually local modifiers or specifications.
Features Based on Modification	
$isPreModified(m)$	If a mention is pre-modified, it tends to be more prominent than unmodified mentions. If a mention is more prominent, it is more likely to have a larger scope.
$headOfRelClause(m)$	Mentions that are the head of a relative clause are usually more prominent and are more likely to have an intermediate or global scope.
Features Based on the Text Structure	
$inSubjPosition(m)$	Mentions in theme position, which is in English often the subject, tend to pick up what has already been mentioned before (Daneš, 1974). Since this is not just the case on the reference-level, but also on the concept-level, the mention in theme position tends to be related to other mentions in the text and tends to have an intermediate or global scope.
$posInSentence(m, q)$	The earlier a mention appears in the sentence in English, the more thematic it is, and the more likely it has an intermediate or global scope.
$focusingAdverb(m)$	Focusing adverbs in the text pattern <*focusing adverb*> <*mention*> – e.g. "particularly *Jack*" – indicate that the mention is thematic and therefore has larger scope.
$modifiesArgument(m)$	A premodifier of a verbal argument is usually more likely to be of local scope.
$passiveBy(m)$	A passive construction – e.g. "the thief was catched by the police" – is a way to reduce the prominency of the agent (e.g. *police*). The agent tends to be of local scope.
$inConjunction(m)$	Conjunctions are often used for exemplifications. Therefore mentions in conjunctions are often less prominent.
$inDepRelPP(m_1, m_2)$ $inDepRelGen(m_1, m_2)$	In NPs with prepositional or genitive modifiers usually at most one part – either the modifying NP or the head – has intermediate or global scope.
$morphoTiesHead(m, q)$	The more frequent the head of a mention appears in the text – also as a derivation, e.g. a verb, according to CatVar (Habash and Dorr, 2003) –, the more prominent it is.
$positionInText(m, q)$	The earlier a mention appears in text, the more likely it is to exhibit global cohesive scope (cf. the hard-to-be-beat lead baseline in summarization (Radev et al., 2003)).

Table 2: Features for cohesive scope distinction. m, m_1, m_2 denote mentions, q a score. The predicates are plugged in the template formula $f8$ in Table 1.

Data set	No. of Mentions	Non-NILs	NILs	Avg. Ambiguity
WP Training	56,372	53,097	3,275	2.31
WP Dev	9,992	9,375	617	2.28
ACE 2005	29,300	27,184	2,116	6.52
ACE 2004	306	257	49	5.04
TAC 2011	2,250	1,124	1,126	6.32
MSNBC	756	629	127	5.29

Table 3: Statistics for data sets.

cept). The average ambiguity of mentions is given by our lexicon (see Section 4.2).

Our system is exclusively trained on the internal hyperlinks in Wikipedia with the advantage that no manual annotation effort is needed. We use 500 articles for training and 100 articles for development (Fahrni and Strube, 2012). Each internal hyperlink is considered as an annotated mention. The pointer to the Wikipedia article serves as the correct concept for this mention and all other candidate concepts we obtain from our lexicon as wrong concepts for this mention.

For the detailed analysis of our approach, we use a version of the ACE 2005 corpus which contains Wikipedia link annotations (Bentivogli et al., 2010). All ACE mentions, both common and proper nouns, are annotated with one or more links to the English Wikipedia or as NILs. If a mention is annotated with more than one link, we consider it as correctly disambiguated if one of the annotated concepts has been chosen by our system. ACE 2005 consists of 597 texts from newswire reports, broadcast news, internet sources and transcribed audio data and contains more annotations than the other data sets we use for comparison.

While ACE 2005 and ACE 2004 (Ratinov et al., 2011) fit our target scenario most (both common and proper nouns are annotated), MSNBC (Cucerzan, 2007) and TAC 2011 (Ji et al., 2011) are only annotated for proper nouns.

4.2 Preprocessing

The training, development and testing data are all preprocessed in the same way. We perform POS tagging, syntactic parsing and named entity recognition using the *Stanford CoreNLP* pipeline[5]. For identifying mentions we extract all noun phrases (excluding discontinuous phrases and determiners) and look them up in our lexicon. Our lexicon and also all other information we obtained from Wikipedia are extracted from the same English Wikipedia dump.[6] The lexicon consists of anchor texts, article titles and redirects.

4.3 Settings

Upper bound: The upper bound shows the maximum performance we can reach given our lexicon and preprocessing. If the correct concept is among the candidate concepts of a mention, it is considered as correct.

First Concept: The first concept baseline is a strong baseline in disambiguation. It chooses for each mention its most frequent concept.

Scope-ignorant (Disambig.): Our previous MLN-based approach for concept disambiguation (Fahrni and Strube, 2012).

Scope-ignorant (Disambig. & Clust.): Our previous MLN-based approach for joint disambiguation and clustering of concepts (Fahrni and Strube, 2012).

Pipeline-based Scope-aware (Disambig.): We compare our joint approach to a pipeline-based one in which the assignment of the cohesive scope is done before disambiguation. The features for the scope assignment and the disambiguation task are exactly the same as in the joint setting and implemented in Markov Logic. The weights for the scope assignment and disambiguation task are learned in a cascaded way. In contrast to the joint approach, the *hasScope(m, s)* predicate is observed during disambiguation.

Joint Scope-aware (Disambig.): This is our approach as described in Section 3 for concept disambiguation. As only local optimization is possible, initialization is crucial. We use the same initialization strategy as for the cascaded approach.

Joint Scope-aware (Disambig. & Clust.): This is our approach as described in Section 3 for disambiguation and clustering of concepts.

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

[6] We use the English Wikipedia dump from Jan. 4, 2012.

4.4 Analysis of Scope-awareness on ACE 2005

In Table 4 we report precision (*P*), recall (*R*) and F-measure (*F*) for non-NILs and NILs for the ACE 2005 data. We also report overall accuracy (*Acc*) (aka micro-average) and calculate significance using a paired t-test.

Differences in the results can be exclusively traced back to differences in the modeling (scope-ignorant vs. scope-aware) and learning (pipeline-based vs. joint). Learning scope-specific models (pipeline-based or joint) significantly improves the result with $p < 0.01$ while using the same features for disambiguation. Scope-aware joint approaches significantly outperform the other corresponding approaches (pipeline-based and scope-ignorant) that use the same features for disambiguation (and clustering) with $p < 0.01$. While the pipeline-based approach suffers from error propagation, the joint approach also benefits from the learning strategy: learning weights for scope distinction can be guided by the training data available for the disambiguation task. Joint disambiguation and clustering of mentions improves the disambiguation results for both the scope-ignorant (Fahrni and Strube, 2012) and the scope-aware approach.

As Table 4 indicates, the gain of the joint scope-aware approach with respect to non-NILs is substantial in both precision and recall. For NILs the recall improves while the precision decreases. This leads to a slightly worse F-Measure for the NILs. As NILs are much rarer than non-NILs in the corpus, the overall accurracy for which we optimize is significantly higher for the scope-aware approaches.

As no gold annotations for cohesive scopes are available, we present statistics on the distribution of induced scopes. Table 5 shows the distribution of the mentions across induced scopes. Mentions with local scope are more frequent than mentions with intermediate scope followed by mentions with global scope. Table 5 compares the overall accurracy of the scope-ignorant joint disambiguation and clustering approach (Fahrni and Strube, 2012) with the accurracy of the corresponding joint scope-aware approach. The joint scope-aware approach improves the disambiguation results for mentions of all three scopes. The biggest gain (2.79) is achieved for mentions with induced global scope. The gain for mentions with local and intermediate scope is 1.27 and 0.3 re-

	Non-NILs			NILs			
	P	**R**	**F**	**P**	**R**	**F**	**Acc**
Upper bound	94.8	91.8	93.3	71.3	100.0	83.3	92.4
First Concept	68.6	70.0	69.3	55.3	40.3	46.6	67.9
Scope-ignorant (Disambig.) (Fahrni & Strube 2012)	77.3	76.0	76.6	44.7	54.2	49.0	74.4
Scope-ignorant (Disambig. & Clust.) (Fahrni & Strube 2012)	76.8	76.9	76.9	50.2	50.0	50.1	74.9
Pipeline-based Scope-aware (Disambig.)	80.1	75.8	77.9	37.3	63.4	47.0	74.9
Joint Scope-aware (Disambig.)	80.1	76.6	78.3	39.2	61.5	47.9	**75.5**
Joint Scope-aware (Disambig. & Clust.)	80.3	77.1	78.6	40.8	62.1	49.3	**76.0**

Table 4: Evaluation on ACE 2005 data

	Scope-ignorant Approach (Disambig. & Clust.) (Fahrni & Strube 2012) (Acc)	**Joint Scope-aware Approach (Disambig. & Clust.) (Acc)**	**Scope Distribution (%)**
Global Scope	73.20	75.99	8.54
Intermediate Scope	76.34	76.64	31.05
Local Scope	75.57	76.84	60.40
Total	75.61	76.71	100.00

Table 5: Evaluation on ACE 2005 data across induced scopes. The accurracy of the two compared systems is slightly higher than in Table 4 as we consider here only mentions that have been recognized by our mention identification strategy. In the evaluation in Table 4 mentions that have not been recognized are considered as wrong.

spectively. A comparison of the learned weights for the different scope-specific models shows that for mentions with local scope the local context has relatively more weight than for mentions with intermediate scope. For mentions with global scope, it is striking that candidiate concepts that are not related to the global context are relatively higher punished than in the other two models.

To obtain some insights on the behaviour of the joint scope-aware approach, we investigate some examples. In a text on the 2004 US elections, the mention *Kerry* in "*Kerry* was the clear winner, but victory was snatched from him" is wrongly disambiguated to KERRY GAA, a branch of the Gaelic football association, by the scope-ignorant approach, because the local context strongly prefers an interpretation in the domain of sports. In the joint scope-aware approach, *Kerry* is assigned global scope, and it is correctly disambiguated to JOHN KERRY, an American politician, as the global relatedness overrules the local context in this model. In another text on U.S. troops in Iraq, the scope-ignorant approach disambiguates *south* in "Monday's advances came one day after British forces in the *south* made their deepest push into Iraq's second largest city" to SOUTHERN UNITED STATES as concepts related to the USA are quite prominent in the text. In the scope-aware approach *south* is considered as being of local scope and is correctly disambiguated as SOUTH. In "we happen to be at a very nice *spot* by the beach where this is a chance for people to get away from cnn coverage" *spot* is disambiguated as SPOT (SATELLITE) in the scope-ignorant approach (misled by CNN), while it has been correctly recognized as NIL by the scope-aware approach in which it is considered as being of intermediate scope. The remaining disambiguation errors can be traced back to (1) scope assignment errors and (2) disambiguation errors (e.g. *Palmisano* (global scope) is disambiguated as SAMUEL J. PALMISANO, but the text refers to a different unknown Palmisano).

4.5 Comparison to State-of-the-art Approaches

Compared to the state-of-the-art for concept and entity disambiguation our approach performs favorably (Table 6). On ACE 2004 (Ratinov et al., 2011) – which contains annotations for common and proper nouns and fits our target scenario most – our scope-aware approach outperforms recent state-of-the-art approaches for concept and entity disambiguation, i.e. Ratinov et al. (2011) and Cheng and Roth (2013). We also ran Ratinov et al.'s (2011) sytem on ACE 2005, but it seems that its mention recognition is not designed for ACE 2005.

We also evaluate our system on the task of entity linking, i.e. the disambiguation of (selected) proper nouns (MSNBC and TAC 2011). Our system fails to beat the best systems, but still

System	ACE 2004	MSNBC	TAC 2011			
	BOC	BOC	Acc	B^3 P	B^3 R	B^3 F1
Ratinov et al. 2011; Cogcomp	77.3	74.9	78.7	75.7	76.5	76.1
Cheng & Roth 2013	85.3	81.2	86.1	82.9	84.5	83.7
Monahan et al. 2011 (Best System at TAC 2011)			86.1	84.4	84.7	84.6
Scope-ignorant (Disambig. & Clust.) (Fahrni & Strube 2012)	83.4	76.5	84.8	82.5	83.0	82.8
Joint Scope-aware (Disambig. & Clust.)	86.3	79.0	85.5	83.6	82.7	83.1

Table 6: Evaluation on various data sets using the respective standard evaluation metrics. *BOC* stands for Bag-of-Concepts. We use the code of Ratinov et al. (2011) to evaluate on ACE 2004 and MSNBC. For TAC 2011, we use the offical evaluation script and report the micro-average (*Acc*) and B^3 scores. Note that for TAC we use three additional disambiguation features – they measure the similarity of the article name to the context – both in the scope-ignorant and the scope-aware approach.

achieves competitive performance without training on TAC data. On all data sets, the joint scope-aware approach consistently outperforms the scope-ignorant approach ceteris paribus.

5 Related Work

Joint approaches have been successful in the past in NLP (e.g. Meza-Ruiz and Riedel (2009)). The idea of augmenting a model with additional latent variables to increase its expressiveness is known as *hidden* or *latent variable learning* (Smith, 2011) and is a promising research direction with successful applications in e.g. syntactic parsing (Petrov et al., 2006), statistical machine translation (Blunsom et al., 2008) and sentiment analysis (Yessenalina et al., 2010; Trivedi and Eisenstein, 2013). For latent variable learning generative approaches (Petrov et al., 2006), large margin methods (Smith, 2011) and conditional log-linear models have been proposed. We focus here on conditional log-linear models due to their flexibility and their previous success for many tasks. Blunsom et al. (2008) for instance use latent variables in the context of discriminative machine translation and model the derivation as a latent variable. Chang et al. (2010) is close to our approach, as their latent variable approach also uses ILP. Poon and Domingos (2008) also use latent variables with Markov Logic, although with a completely different aim, i.e. for unsupervised coreference resolution.

Most approaches that use Wikipedia as a resource for disambiguation focus on named entities (Bunescu and Paşca, 2006; Cucerzan, 2007; Dredze et al., 2010; Ji and Grishman, 2011; Hachey et al., 2013; Hoffart et al., 2011), while only a few disambiguate common and proper nouns like us (Csomai and Mihalcea, 2008; Milne and Witten, 2008; Zhou et al., 2010; Ratinov et al., 2011; Cheng and Roth, 2013). We build upon our previous Markov Logic based approach for joint concept disambiguation and clustering (Fahrni and Strube, 2012). In contrast to us, most approaches for lexical disambiguation use either one model for all mentions (Milne and Witten, 2008; Ratinov et al., 2011) or a separate model for each mention or concept which requires a lot of training data (e.g. Bryl et al. (2010)). Only a few approaches try to learn specific models for groups of mentions, although none of them is discourse-motivated as ours: Mihalcea and Csomai (2005) learn a specific model for each POS, Ando (2006) uses alternating structure optimization to simultantanously learn a number of WSD problems and Dhillon and Ungar (2009) improve feature selection for WSD by integrating knowledge from similar words.

6 Conclusions

In this paper, we discuss the relationship between cohesion and concept disambiguation and propose a cohesive scope-aware disambiguation approach. We distinguish between three different cohesive scopes (local, intermediate and global) and model the scope assignment and the disambiguation jointly using latent variables in the framework of MLN. The joint scope-aware approach significantly improves over both a state-of-the-art and a pipeline-based approach using the same features for the disambiguation task.

For future work, we are planning to investigate the relation between discourse structure and cohesive scope more deeply and to integrate scope-specific disambiguation features.

Acknowledgments

We would like to thank Sebastian Martschat for his valuable comments. This work has been partially funded by the Klaus Tschira Foundation.

References

Rie Kubota Ando. 2006. Applying alternating structure optimization to word sense disambiguation. In *Proceedings of the 10th Conference on Computational Natural Language Learning,* New York, N.Y., USA, 8–9 June 2006, pages 77–84.

Nicholas Asher and Alex Lascarides. 1995. Lexical disambiguation in a discourse context. *Journal of Semantics*, 12(1):69–108.

Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta, and Kateryna Tymoshenko. 2010. Extending English ACE 2005 corpus annotation with ground-truth links to Wikipedia. In *Proceedings of the 2nd Workshop on The People's Web: Colloboratively Constructed Semantic Resources,* Beijing, China, 28 August 2010, pages 19–27.

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* Columbus, Ohio, 15–20 June 2008, pages 200–208.

Volha Bryl, Claudio Giuliano, Luciano Serafini, and Kateryna Tymoshenko. 2010. Supporting natural language processing with background knowledge: Coreference resolution case. In *Proceedings of the 9th International Semantic Web Conference, Revised Selected Papers, Part I,* Shanghai, China, 7-11 November 2010, pages 80–95.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006, pages 9–16.

Ming-Wei Chang, Vivek Srikumar, Dan Goldwasser, and Dan Roth. 2010. Structured output learning with indirect supervision. In *Proceedings of the 27th International Conference on Machine Learning,* Haifa, Israel, 21–24 June 2010, pages 199–206.

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.

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 Language Learning,* Prague, Czech Republic, 28–30 June 2007, pages 708–716.

František Daneš. 1974. Functional sentence perspective and the organization of the text. In F. Daneš, editor, *Papers on Functional Sentence Perspective*, pages 106–128. Prague: Academia.

Paramveer S. Dhillon and Lyle H. Ungar. 2009. Transfer learning, feature selection and word sense disambguation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers,* Singapore, 2–7 August 2009, pages 257–260.

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, Benjamin Heinzerling, Thierry Göckel, and Michael Strube. 2014. HITS' monolingual and cross-lingual entity linking system at TAC 2013. In *Proceedings of the Text Analysis Conference,* National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 18–19 November 2013.

Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for English. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Edmonton, Alberta, Canada, 27 May –1 June 2003, pages 17–23.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with Wikipedia. *Artificial Intelligence*, 194:130–150.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 362–370.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. London, U.K.: Longman.

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, Edinburgh, Scotland, U.K., 27–29 July 2011, pages 782–792.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, Oreg., 19–24 June 2011, pages 1148–1158.

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.

Daniel Lowd and Pedro Domingos. 2007. Efficient weight learning for Markov logic networks. In *Proceedings of the 11th European Conference on Principles and Practices of Knowledge Discovery in Databases,* Warsaw, Poland, 17–21 September 2007, pages 200–211.

Ivan Meza-Ruiz and Sebastian Riedel. 2009. Jointly identifying predicates, arguments and senses using Markov logic. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 155–163.

Rada Mihalcea and Andras Csomai. 2005. SenseLearner: Word sense disambiguation for all words in unrestricted text. In *Proceedings of the Interactive Poster and Demonstrations Sessions at the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June 2005, pages 53–56.

Rada Mihalcea. 2006. Knowledge-based methods for WSD. In E. Agirre and P.G. Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, pages 107–131. Springer, Heidelberg, Germany.

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.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Rani Nelken and Stuart Shieber. 2006. Lexical chaining and word-sense-disambiguation. Technical Report TR-06-07, Computer Science Group, Harvard University, Cambridge, Mass.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. LDC2011T07.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* Sydney, Australia, 17–21 July 2006, pages 433–440.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 650–659.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pages 375–382.

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.

Noah A. Smith. 2011. *Linguistic Structure Prediction*. Morgan & Claypool Publishers.

Rakshit Trivedi and Jacob Eisenstein. 2013. Discourse connectors for latent subjectivity in sentiment analysis. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Atlanta, Georgia, 9–14 June 2013, pages 808–813.

Ainur Yessenalina, Yejin Choi, and Claire Cardie. 2010. Automatically generating annotator rationales to improve sentiment classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pages 336–341.

Yiping Zhou, Lan Nie, Omid Rouhani-Kalleh, Flavian Vasile, and Scott Gaffney. 2010. Resolving surface forms to Wikipedia topics. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pages 1335–1343.