Trust, but Verify!
Better Entity Linking through Automatic Verification

Benjamin Heinzerling*
AIPHES
Heidelberg Institute for
Theoretical Studies
benjamin.heinzerling@h-its.org

Michael Strube
Heidelberg Institute for
Theoretical Studies
michael.strube@h-its.org

Chin-Yew Lin
Microsoft Research
cyl@microsoft.com

Abstract

We introduce automatic verification as a post-processing step for entity linking (EL). The proposed method *trusts* EL system results collectively, by assuming entity mentions are mostly linked correctly, in order to create a semantic profile of the given text using geospatial and temporal information, as well as fine-grained entity types. This profile is then used to automatically *verify* each linked mention individually, i.e., to predict whether it has been linked correctly or not. Verification allows leveraging a rich set of global and pairwise features that would be prohibitively expensive for EL systems employing global inference. Evaluation shows consistent improvements across datasets and systems. In particular, when applied to state-of-the-art systems, our method yields an absolute improvement in linking performance of up to 1.7 $F1$ on AIDA/CoNLL'03 and up to 2.4 $F1$ on the English TAC KBP 2015 TEDL dataset.

1 Introduction

Entity linking (EL) is the task of automatically linking mentions of entities such as persons, locations, or organizations to their corresponding entry in a knowledge base (KB). The task is generally approached by generating a set of candidate entities[1] for a given mention and then ranking those candidates. Approaches differ in whether they rank a mention's candidates independently of the candidates of other mentions ("local inference") or whether they rank all candidates of all mentions simultaneously by incorporating a global coherence measure into the optimization goal ("global inference").

While linguistically well-founded in the concept of lexical cohesion (Halliday and Hasan, 1976), global inference approaches (Kulkarni et al., 2009; Hoffart et al., 2011a) do not scale well with number of mentions and number of candidate entities. In contrast, local approaches do not suffer from scalability issues, since they only optimize the similarity between mention context and candidate KB entry text (Bunescu and Paşca, 2006; Cucerzan, 2007), usually also including a popularity prior[2] (Milne and Witten, 2008; Spitkovsky and Chang, 2012). Recent local approaches achieve state-of-the-art results by using convolutional neural networks to capture similarity at multiple context sizes (Francis-Landau et al., 2016), but, by definition, fail to take global coherence into account.

To avoid the trade-off between the efficiency of local inference on the one hand and the coherence benefits of global inference on the other, we propose a two-stage approach: In the first stage, candidate entities are ranked by a fast, local inference-based EL system. In the second stage these results are used to create a semantic profile of the given text, derived from rich data the KB contains about the top-ranked candidates. Since the linking precision of current EL systems is relatively high, we trust that this profile is reasonably accurate and leverage it to measure the cohesive strength between a given candidate entity and the other linked entities mentioned in the text. We then automatically verify the first stage results by classifying entity links as correct if they display high coherence, and as wrong if there are only weak or no cohesive

*The majority of this work was done during an internship at Microsoft Research Asia.

[1] We use *entity* to refer to both real-word entities and to their corresponding entries in the KB.

[2] Also referred to as *commonness prior* by some authors.

ties to the semantic profile. Verification results can be used in at least three ways:

1. To increase linking precision by filtering out all entity links classified as wrong;

2. To rerank candidate entities by the class probability estimated by the verifier, i.e., prefer candidates that were predicted as correct with higher probability; or

3. To employ a more sophisticated EL system to re-link all entity links classified as wrong, using the entity links deemed correct as additional context.

In this work we investigate options 1. and 2., and make the following contributions:

- We propose automatic verification as a post-processing step for EL systems;

- We propose global coherence features based on notions of entity type coherence, geographic coherence, and temporal coherence;

- We show how these novel features, as well as features developed in prior work, can be used to verify EL results; and

- We show that automatic verification consistently improves linking performance in an evaluation across two datasets and seven different EL systems.

2 Method

We cast entity linking verification as a supervised classification task. Given EL system output on a training set with gold standard linked entity annotations, we extract global, pairwise, and local features and train a classifier to predict whether a given mention has been linked correctly by the EL system.

In the standard EL setting, global inference is an NP-hard problem, since all combinations of all candidate entities of all mentions are considered simultaneously. In our proposed automatic verification setting, however, taking only the top candidate entities into account allows us to employ knowledge-rich, global coherence features that would be prohibitively expensive otherwise.

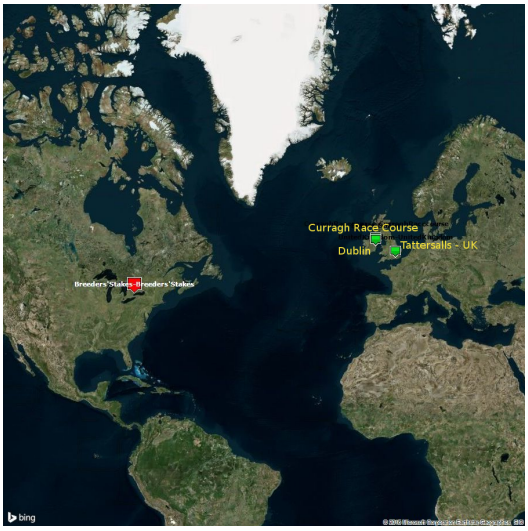

Figure 1: Example showing a geographical outlier: Breeder' Stakes (red, in Canada) and contextual entities located in Ireland and the UK (green).

2.1 Aspects of Global Coherence

Global coherence captures how well a candidate entity fits into the overall semantic profile of a text. Current global inference approaches optimize a single coherence measure, most commonly a measure of general semantic relatedness such as the Milne-Witten distance (Milne and Witten, 2008), or keyphrase overlap relatedness (KORE) (Hoffart et al., 2012).

In contrast, verification allows employing many global coherence features, which we categorize according to four aspects of coherence: geographical coherence and temporal coherence, which to our knowledge have not been used before in EL, as well as entity type coherence and the general semantic relatedness mentioned above.

2.1.1 Geographic Coherence

Entities mentioned in a text tend to be geographically close or clustered around very few locations. We use this observation to identify geographic outliers as potential entity linking mistakes.

For example, consider the mention *Breeders Stakes* in the following excerpt (CoNLL 1112testa):

> DUBLIN 1996-08-31 Result of the Tattersalls Breeders Stakes , a race for two-year-olds run over six furlongs at The Curragh ...

DUBLIN, *Tattersalls* (a company doing business in the UK and Ireland), and *The Curragh* (a

Predicate
:location.location.geolocation
:organization.organization.geographic_scope
:time.event.locations
:sports.sports_team.location
:organization.organization.headquarters

Table 1: Freebase predicates for querying geo-coordinates of locations, geo-political entities, and organizations.

Predicate
:people.person.date_of_birth
:organization.organization.date_founded
:sports.sports_team.founded
:location.dated_location.date_founded
:time.event.start_date
:film.film.initial_release_date
:music.album.release_date
:music.release.release_date
:architecture.structure.construction_started
:architecture.structure.opened
:people.deceased_person.date_of_death
:location.dated_location.date_dissolved
:time.event.end_date
:business.defunct_company.ceased_operations
:architecture.structure.closed

Table 2: Freebase predicates for querying the *begin* (top) and *end* (bottom) of an entity's temporal range.

horse race track in Ireland) clearly situate the text in Ireland (cf. Figure 1). However, some current EL systems link *Breeder Stakes* to the Wikipedia article about the Canadian horse race of the same name, since the Irish race does not have a Wikipedia article and other evidence[3] suggests a strong match.

We aim to identify these kinds of errors by first querying locations (Table 1) of all linked mentions in the document, and then performing geographic outlier detection[4]. This yields a binary feature indicating whether a candidate entity is a geographic outlier or not.

Since outliers are rare and hence the resulting features sparse, we also also add a feature for the average geographic distance $\bar{d}(d, E)$ of a candidate entity e to all other entities in document D:

$$\bar{d}(e, D) = \frac{\sum_{e' \in D \setminus e} d(e, e')}{|D| - 1}$$

where $d(e, e')$ is the geographic distance between entities e and e', and $|D|$ is the number of entities mentioned in D. This feature is based on the intuition that a candidate entity which is geographically closer to other entities is more likely to be correct than a distant one.

Geographic scope varies across documents. For example, entities mentioned in a text about world politics will be geographically more distant than entities in a text about a local business). As a scale-invariant distance measure $s(e, D)$, we divide the average distance $\bar{d}(e, D)$ by the average distance between all other entities:

$$s(e, D) = \bar{d}(e, D) / \frac{\sum_{e', e'' \in D \setminus e} d(e', e'')}{|e', e'' \in D \setminus e|}$$

2.1.2 Temporal Coherence

Applying the notion of coherence to the temporal dimension, we observe that entities mentioned in a text tend to be temporally close or clustered around a few points in time.

Entities are associated with temporal ranges with a *begin*, i.e. the point in time at which the entity comes into existence, and an *end*, i.e. the point in time at which the entity ceases to exists. Using the same approach as in geographical outlier detection, we perform temporal outlier detection on all *begin* and *end* times associated with linked entities in the given text, and declare a candidate entity as an temporal outlier if both its *begin* and *end* were detected as outliers.

Since temporal outliers are rare, we also add a feature aiming to capture temporal proximity and distance in a softer fashion with higher coverage; by calculating the total overlap $T(e, D)$ between the temporal range $t(e)$ of a candidate entity e, and the known temporal ranges of all other linked entities in the document D:

$$T(e, D) = \sum_{e' \in D \setminus e} \left| t(e) \cap t(e') \right|$$

where $|t(e) \cap t(e')|$ is the length of the overlap between the temporal ranges of entities e and e'.[5]

Analogously to the geographic distance feature, we take temporal proximity, i.e. a large overlap with other temporal ranges, as evidence for a correctly linked entity, and temporal distance, i.e. only small or no overlap with other temporal

[3] Specifically, high context-similarity due to the appositive *race*, and an almost perfect string match between mention and Wikipedia title.

[4] We use an ensemble of standard outlier detection algorithms provided by the ELKI clustering toolkit (Achtert et al., 2011).

[5] We also extract this feature normalized by the number of entity mentions in the document, but did not see any effect. This is likely due to little variation in the number of entities per document for which the KB contains temporal information.

ranges, as evidence for a linking mistake. Temporal ranges are queried from the KB using the predicates shown in Table 2.

The final feature using temporal information checks whether an entity's temporal ranges contains the document's creation date. This feature is based on the intuition that, especially in the news genre, an existing entity is more likely to be mentioned than an entity that has already ceased to exist or did not exists at the time of writing. The document creation date is either trivially obtained if metadata is present, or heuristically by using the first date found in the document text by the HeidelTime temporal tagger (Strötgen and Gertz, 2010).

2.1.3 Entity Type Coherence

Frequency statistics of the types of entities mentioned in a text are an indicator of what the text is about. For example, looking at the entity type distribution shown in Table 3, we can tell that the corresponding text appears to be about rugby teams. Unlike other methods for representing the "aboutness" of a text, such as topic models, entity type statistics are grounded in the KB, thus offering a simple method of measuring the relatedness between entities in terms of their types via the similarity of their type distributions.

Specifically, we model entity type coherence between a given candidate entity e and all other linked entities in document D as the cosine similarity of the respective type distributions. Type frequencies are TF-IDF weighted, in order to discount frequent types (e.g. `:base:tagit.concept`) and give more importance to salient types occurring in the document (e.g. `:base.rugby.rugby_club`):

$$coh_{type}(e, D) = sim(types(e), tfidf(types(D)))$$

where sim is the cosine similarity, $types(e)$ a binary vector indicating the types of entity e, and $types(D)$ a vector whose entries are occurrence counts of entity types in document D, which are weighted by $tfidf$.

2.1.4 Semantic Relatedness

Measures of generic semantic relatedness are a standard feature in global inference systems. We add features for the average and maximum semantic relatedness $SemRel(e, D)$ of a candidate entity e with respect to all other entities e' mentioned in document D, using two semantic relatedness measures:

$$SemRel_{max}(e, D) = max_{e' \in D \setminus e} SemDist(e, e')$$

$$SemRel_{avg}(e, D) = avg_{e' \in D \setminus e} SemDist(e, e')$$

where max and avg are the maximum and average operators. $SemDist$ denotes either the Milne-Witten Distance (Milne and Witten, 2008), which defines relatedness of Wikipedia entries in terms of shared incoming article links, or the Normalized Freebase Distance (Godin et al., 2014), an adaptation of the Milne-Witten Distance to Freebase entities.

2.2 Pairwise Features

Semantic relation: Given a pair consisting of a candidate entities and an entity mention in its context, we add a feature encoding whether a (and if yes which) semantic relation exists between the two entities. We add different features depending on the type of context in which the entity pair occurs: in the same sentence, within a fixed token window, and within the same noun phrase. For example, in the noun phrase *German Chancellor Angela Merkel*, we find a `wasBornIn` and a `isLeaderOf` relation between YAGO entities ANGELA_MERKEL[6] and GERMANY. We expect this feature to be sparse, but strong evidence for both arguments of the identified relation being linked correctly. We record the relation type, as some relations tend to be more informative than others, e.g., the `playsFor` relation, which holds between players and sports teams,

[6] In this work, SMALL_CAPS denote both real-world entities and their corresponding entries in the knowledge base.

TF-IDF	Count	Type
1115.67	2	`:base.rugby.rugby_club`
243.62	3	`:organization.organization`
231.76	2	`:base.schemastaging.sports_team_extra`
183.49	2	`:sports.sports_team`
56.34	2	`:base.tagit.concept`

Table 3: Entity type distribution in a document about rugby, sorted by type TF-IDF.

should provide stronger evidence than the less specific `isCitizenOf` relation, which holds between citizens and countries.

Person name consistency: Having observed that some local inference systems tend to make the mistake of linking a full name mention (e.g. "John Smith") to one entity, and a coreferent surname-only mention ("Smith") to a different one, we add a binary feature that indicates whether a candidate entity assigned to a partial person name mention agrees with its unambiguous full name antecedent.

2.3 Local Features

Since the global and pairwise features do not have high enough coverage to provide evidence for all linked candidate entities, we employ local features that are devised to capture similarity between a candidate entity and its textual context. As these features are commonly used in EL systems, we only give brief descriptions for completeness.

Popularity prior: The prior probability of the candidate entity given its mention, obtained from the CrossWikis dictionary (Spitkovsky and Chang, 2012). This feature aims to cover unambiguous and almost unambiguous mentions.

Entity type agreement: A binary feature indicating whether the candidate entity type, as found in the KB agrees with the named entity type, as determined by the NER system during preprocessing.

Keyphrase match: Knowledge bases contain various sources of key phrases, such as labels and aliases of semantic types, or salient noun phrases in description texts, e.g., noun phrases occurring in the first, defining sentence of a Wikipedia article. We add a binary feature indicating whether a known keyphrase occurs in the context of a given candidate entity.

Demonym match: This binary feature indicates whether a mention is a demonym of its linked entity, e.g., the mention text *French* is a demonym match for the entity FRANCE.

Mention-entity string match: Finally, we extract features from the string similarity between a mention and the known labels and aliases of a candidate entity. The similarity measures include exact match, case-insensitive match, head match, match with stop words filtered, fuzzy string match, Levenshtein distance, and abbreviation pattern matches, as well as different combinations of these.

Dataset	CoNLL	TAC15
Entity Type	99.2	98.5
Geographic	62.9	41.5
Temporal	87.6	79.4

Table 4: KB coverage of our proposed global coherence features. Shown are the percentages of in-KB mentions in each dataset for which the KB (YAGO or Freebase) contains the required information for each coherence feature set.

3 Experiments

We evaluate our automatic verification method by applying it to the entity linking results produced by seven systems on two standard datasets: CoNLL, which consists of 1393 Reuters news articles annotated with Wikipedia links by Hoffart et al. (2011a) and TAC15, which comprises 315 news articles and discussion forum texts annotated with Freebase links for the TAC KBP 2015 TEDL shared task (Ji et al., 2015).

The KB coverage for each of our proposed global coherence features on these two datasets is shown in Table 4. YAGO and Freebase contain entity type information for almost all in-KB entities mentioned in the two datasets. Geographic data is available for 62.9 percent on CoNLL, but only for 41.5 percent of entities mentioned in TAC15. This difference is likely due to the large fraction of documents from the sports genre in CoNLL. These documents include match result tables mentioning a large number of sports teams, which can be easily located via their cities and stadiums. Temporal information is present for most entities.

Our evaluation uses results of the following EL systems:

AIDA (Hoffart et al., 2011a): This system globally optimizes a graph-based model incorporating three factors: a popularity prior, the context similarity of mention and candidate entity, and coherence modeled via general semantic relatedness measures. We use the AIDA system output on the CoNLL dataset as provided by the Wikilinks project.[7]

SPOTL (Daiber et al., 2013): DBpedia Spotlight is a local inference system. We use results obtained from the Spotlight webservice.[8]

[7] https://github.com/wikilinks/conll03_nel_eval

[8] https://github.com/dbpedia-spotlight/

FL (Francis-Landau et al., 2016): This local inference system models mention and entity context with a convolutional neural network (CNN). The CNN captures semantic similarity of a given mention's context at different granularities (small context window, paragraph, document) and the entity context derived from the entity's Wikipedia page.
PH (Pershina et al., 2015): This global inference system applies Personal PageRank to a graph whose nodes represent candidate entities and whose edges indicate if a link between the corresponding Wikipedia articles exists. PH achieves the best CoNLL performance among the systems in our evaluation.
TAC-1 (Heinzerling and Strube, 2015): This system uses local and pairwise inference in an easy-first, incremental rule-based approach. Features are based on popularity priors, contextual occurrence of keywords, entity type, and relational evidence.
TAC-2 (Sil et al., 2015): This system employs a global inference approach which partitions a document into sets of mentions that appear near each other. The partitioning is motivated by the intuition that a given mention's immediate context provides the most salient information for disambiguation, and drastically reduces the search space during global optimization.
TAC-3 (Dai et al., 2015): This local inference system models mentions and entity context with a CNN and word embeddings.

The systems were chosen for their popularity (AIDA, SL), performance on CoNLL (FL, PH), and performance on TAC15 (TAC systems). Unless stated otherwise, we use system output provided by authors for CoNLL systems, and provided by the workshop organizers for TAC15 systems.[9] Our evaluation does not include (Globerson et al., 2016) and (Yamada et al., 2016), who report better performance on CoNLL than PH, but were unable to make system output available.

3.1 Setup and Implementation Details

Feature extraction is implemented as a UIMA pipeline (Ferrucci and Lally, 2004); using the Stanford CoreNLP (Manning et al., 2014) UIMA components provided by DKPro (Eckart de Castilho and Gurevych, 2014) for text segmentation, POS tagging, and named entity recognition; DKPro WSD (Miller et al., 2013) for modeling entity mentions and links, and using Freebase (Bollacker et al., 2008) and YAGO (Hoffart et al., 2011a) as knowledge bases.

After feature extraction, we train a random forest classifier[10] for each dataset, one using FL system results for the CoNLL development set (216 documents) and one using TAC-1 results for the TAC15 training set (168 documents).

For evaluation, we apply the verifier trained on FL CoNLL development results to the test set results of the FL and AIDA systems, and a verifier trained on PH training data to the PH test set results. For the test set output of TAC systems 1-3 we apply the verifier trained on the TAC15 training set output of TAC-1.

As metric we use `strong_link_match` as implemented by the Wikilinks project for the CoNLL dataset, and the official NIST scorer (Hachey et al., 2014) for TAC15. This metric measures precision, recall, and $F1$ of matching entity links and mention spans.

3.2 Results and Discussion

Evaluation results are shown in Table 5. Our method improves the linking performance of all evaluated EL systems. The impact is most noticeable for the systems that only use local and pairwise inference, namely FL (+1.9 $F1$), TAC-1 (+2.4 $F1$), TAC-3 (+1.1 $F1$). The improved TAC-1 result ($68.1F1$) is the best published linking score on the TAC15 dataset.

Improvements are smaller for the global inference systems, AIDA, HP, and TAC-2. In contrast to Ratinov et al. (2011), who report only a very small increase in linking performance when incorporating global features into a local inference-based system, our results indicate that global features are useful and lead to considerable improvements.

As expected, improvements are caused by increased precision, due to filtering out likely linking mistakes. The fact that this increase is not accompanied by a commensurate decrease in recall, shows that our method predicts wrong linking decisions with high accuracy.

On TAC15, we observe considerable improvements in linking precision of up to 10.4 percent.

`dbpedia-spotlight/wiki/Web-service`

[9] `http://www.nist.gov/tac/2015/KBP/data.html`

[10] Various other classifiers we tried, e.g. neural networks, showed no better performance during cross-validation on development sets.

Dataset	System	Baseline			After verification			Δ		
		Prec	Rec	$F1$	Prec	Rec	$F1$	Prec	Rec	F1
CoNLL	AIDA	83.2	**83.6**	83.4	**86.0**	82.3	**84.1**	+2.8	-1.3	+0.7
	SPOTL	85.5	**80.5**	82.9	**93.0**	77.6	**84.6**	+7.5	-2.9	+1.7
	FL	85.3	**85.2**	85.2	**89.2**	84.7	**86.9**	+4.0	-0.5	+1.7
	PH	90.5	**90.5**	90.5	**93.2**	89.1	**91.1**	+2.7	-1.4	+0.6
TAC15	TAC-1	71.2	**61.1**	65.8	**81.6**	58.6	**68.2**	+10.4	-2.5	+2.4
	TAC-2	71.4	**57.9**	63.9	**81.2**	53.3	**64.4**	+9.8	-4.6	+0.5
	TAC-3	68.0	**55.6**	61.1	**77.6**	52.0	**62.2**	+9.6	-3.2	+1.1

Table 5: Results on CoNLL and TAC15 test sets. *Baseline* shows performance of the original systems, *After verification* shows performance after application of our automatic verification method, and Δ shows the corresponding change. Bold font indicates best results for each metric and system.

On CoNLL, the precision increase is less pronounced, arguably owing to the already higher baseline precision, which leaves less room for improvement. Since EL is usually performed as part of a larger task, such as knowledge base completion, search, or as part of a more comprehensive entity analysis system (Durrett and Klein, 2014), good precision is highly desirable in order to minimize error propagation to other system components and downstream applications.

3.3 Candidate Reranking

We resort to the binary decision of either retaining or removing an entity linked by an EL system if no candidate entities and no meaningful confidence scores are available. This is the case for the output of many EL systems, such as the systems participating in the TAC KBP TEDL 2015 challenge.

In case the EL system outputs not only the top-ranked candidate entity, but also lower-ranked ones, we can apply our verification method to all candidates and rerank them according to their probability of being correct. For example, if the EL system linked a mention to candidate entity e_1 over candidate e_2, but verification assigns a higher probability of being correct to e_2, we rerank e_2 over e_1. Since we assume that the document's semantic profile derived from EL results is sufficiently accurate, we do not recreate it after reranking a candidate.

Reranking the candidate entities produced by the FL system on the CoNLL test set, this achieves a similar increase in $F1$, but with a different precision-recall trade-off (Table 6): We observe highest precision at the cost of a lower recall for filtering, while reranking increases both precision and recall.

System	Prec	Rec	F1
FL baseline	85.3	85.2	85.2
FL filter	**89.2**	84.7	**86.9**
FL rerank	87.9	**85.6**	86.7

Table 6: Comparison of filtering and candidate entity reranking performance on the CoNLL test set.

3.4 Ablation Study

We conduct an ablation study to assess the impact of the proposed global coherence features on prediction performance. Applying backward elimination (John et al., 1994), we iteratively remove one feature set and successively eliminate the feature set with the largest impact (Figure 2).

Surprisingly, the string similarity features have a large effect across all three systems. This suggests that current systems do not optimally utilize string similarity when selecting and ranking candidate entities for a given mention.

Our proposed global coherence features are among the top features for all systems. This contradicts prior findings by Ratinov et al. (2011) and shows that global coherence has a considerable impact on EL performance. We believe that this is due to our proposed coherence features being more informative than the generic semantic relatedness measures used in prior work. While ablation indeed shows a relatively low importance of semantic relatedness features (cf. SemRel in Figure 2), further research is required to test this hypothesis.

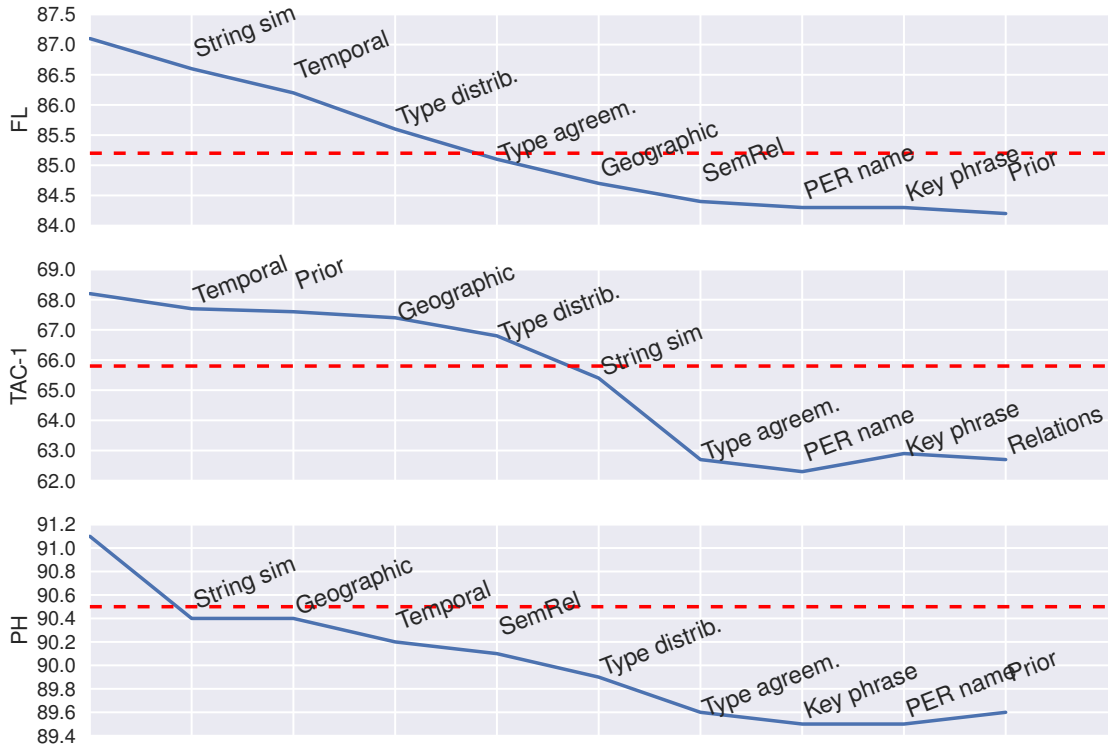

Figure 2: Feature set ablations for the FL, TAC-1, and PH systems. The solid blue lines show the performance impact in terms of `strong_link_match` $F1$ incurred from eliminating feature sets. The red dashed line indicates baseline performance without verification.

3.5 Automatic Verification on Noisy Text

The TAC15 dataset consists of different text genres: clean newswire articles, and noisy discussion forum threads. Analysis of verification performance on these two genres reveals that verification has the biggest impact on noisy text (Table 7, bottom), while the improvement is smaller for two systems on clean text, and even slightly negative for one system, namely the global inference system TAC-2 (Table 7, top).

4 Related Work

Global coherence has been successfully employed for EL in a number of seminal works (Kulkarni et al., 2009; Hoffart et al., 2011b; Han et al., 2011), and more recently by Moro et al. (2014), Pershina et al. (2015), and Globerson et al. (2016), among others. These approaches maximize global coherence based on a general notion of semantic relatedness, while considering a fixed number of candidate entities for each mentions. Our approach differs from these in in two regards. Firstly, we introduce specific aspects of coherence, namely entity type coherence, geographic coherence, and temporal coherence. While these aspects are limited to certain entities, such as entities with a clearly defined location and temporal range, our experiments showed that features based on these notions of coherence are useful on the types of texts found in common datasets. Secondly, in our verification setting, these rich coherence measures can be efficiently incorporated since their computation is linear in the number of entities mentioned in a document, while they would be prohibitively expensive in the global inference EL setting.

Entity types have been used in prior work. Cucerzan (2007) maximizes the agreement of Wikipedia categories associated with candidate entities. Due to intractability of the resulting global optimization problem, the agreement of the candidate entities for a given mention is maximized with respect to all categories of all candidate entites of all other mentions, and hence includes many wrong categories. Our approach is more precise, since verification allows using only the types of the top-ranked candidate entities. Sil and Yates (2013) also employ entity types, but only maximize type agreement of entity mentions in a small context window. In contrast, our ap-

Genre	System	Baseline			After verification			Δ		
		Prec	Rec	$F1$	Prec	Rec	$F1$	Prec	Rec	F1
News	TAC-1	66.5	**60.3**	63.2	**75.8**	57.0	**65.0**	9.3	-3.3	1.8
	TAC-2	69.7	**59.9**	**64.4**	**79.3**	53.9	64.2	9.6	-6.0	-0.2
	TAC-3	63.0	**59.1**	61.0	**71.3**	54.3	**61.7**	8.3	-4.8	0.7
Forum	TAC-1	76.0	**61.8**	68.1	**87.4**	60.0	**71.2**	11.4	-1.8	3.1
	TAC-2	73.1	**56.1**	63.5	**83.0**	52.8	**64.6**	9.9	-3.3	1.1
	TAC-3	73.8	**52.4**	61.3	**84.7**	49.9	**62.8**	10.9	-2.5	1.5

Table 7: Verification on different text genres. See caption of Table 5 for details.

proach uses global context and hence allows capturing long-distance relations.

Post-processing of EL system output has been approached as an ensembling task (Rajani and Mooney, 2016). In this setting, a meta-classifier combines the output of different EL systems on a given dataset, taking into account features such as system confidence scores, past system performance, and number of systems agreeing with a given decision. Our approach differs from ensembling, since we post-process the output of a single system, using rich semantic features. In contrast, ensembling requires multiple system outputs and relies on meta-information about system performance and decision confidence. Combining these two post-processing methods is an interesting problem for future work and could lead to further improvements, since the two methods rely on different types of information.

5 Conclusions and Future Work

We have introduced automatic verification as a post-processing step for entity linking (EL). Our method uses the output of an existing EL system to create a semantic profile of the given text using entity types, as well as geographic and temporal information. Due to the high precision achieved by state-of-the-art EL systems, this profile is a sufficiently accurate representation of the text's main topic, and further situates the text temporally and geographically This profile is then used to automatically verify each linked mention individually, i.e., to predict whether it has been linked correctly or not. Verification allows leveraging a rich set of global and pairwise features that would be prohibitively expensive for EL systems employing global inference. Evaluation showed consistent improvements when applying our method to seven different EL systems on two different datasets.

Our main goal in future work is the better integration of our approach with existing EL systems. Most notably, some EL systems produce meaningful confidence scores, which we currently disregard. We expect further improvements from incorporating various confidence measures into the verification process. Automatic verification could also be used in an easy-first setting to identify likely correct decisions made by a fast and simple EL system, and then perform the remaining decisions with a more sophisticated system. Since our features make use of coreference information in the form of person name agreement, as well as entity types, another line of future research is expanding our proposed entity linking verification method to entity analysis (Durrett and Klein, 2014), which models entity linking, coreference, and entity typing as a joint task.

Acknowledgments

We thank Matthew Francis-Landau, Maria Pershina, as well as the TAC KBP 2015 organizers for providing system output, and the anonymous reviewers for providing helpful feedback. This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1.

References

Elke Achtert, Ahmed Hettab, Hans-Peter Kriegel, Erich Schubert, and Arthur Zimek. 2011. Spatial outlier detection: Data, algorithms, visualizations. In *Advances in Spatial and Temporal Databases - 12th International Symposium, SSTD 2011, Minneapolis, MN, USA, August 24-26, 2011, Proceedings*, pages 512–516.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim

Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data,* Vancouver, B.C., Canada, 10–12 June 2008, pages 1247–1250.

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.

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.

Hongliang Dai, Siliang Tang, Fei Wu, Zewu Ma, and Yueting Zhuang. 2015. The ZJU-EDL system for entity discovery and linking at TAC KBP 2015. In *Proceedings of the Eighth Text Analysis Conference*, Gaithersburg, MD, USA. National Institute of Standards and Technology.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, pages 121–124, Graz, Austria.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association of Computational Linguistics*, 2:477–490.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

David A. Ferrucci and Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3):327–348.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261, San Diego, California, June. Association for Computational Linguistics.

Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringaard, and Fernando Pereira. 2016. Collective entity resolution with multi-focal attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Berlin, Germany, August. Association for Computational Linguistics.

Fréderic Godin, Tom De Nies, Christian Beecks, Laurens De Vocht, Wesley De Neve, Erik Mannens, Thomas Seidl, and Rik Van de Walle. 2014. The normalized freebase distance. In *The Semantic Web: ESWC 2014 Satellite Events*, pages 218–221. Springer, Anissaras, Crete, Greece.

Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and easy entity evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* Baltimore, Md., 22–27 June 2014, pages 464–469.

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

Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: A graph-based method. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Beijing, China, 25–29 July 2011, pages 765–774.

Benjamin Heinzerling and Michael Strube. 2015. Visual error analysis for entity linking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* Beijing, China, 26–31 July 2015, pages 37–42.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. 2011a. YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th World Wide Web Conference,* Hyderabad, India, 28 March – 1 April, 2011, pages 229–232.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011b. 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.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the ACM 21st Conference on Information and Knowledge Management,* Maui, Hawaii, USA, 29 October – 2 November 2010, pages 545–554.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP 2015: Tri-lingual entity discovery and linking. In *Proceedings of the Eighth Text Analysis Conference*, Gaithersburg, MD,

USA. National Institute of Standards and Technology.

George H. John, Ron Kohavi, and Karl Pfleger. 1994. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*, pages 121–129.

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 International Conference on Knowledge Discovery and Data Mining,* Paris, France, 28 June – 1 July 2009, pages 457–466.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

Tristan Miller, Nicolai Erbs, Hans-Peter Zorn, Torsten Zesch, and Iryna Gurevych. 2013. Dkpro wsd: A generalized uima-based framework for word sense disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Sofia, Bulgaria, August. Association for Computational Linguistics.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08,* Chicago, Ill., 13 July 2008, pages 25–30.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized page rank for named entity disambiguation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 238–243, Denver, Colorado, May–June. Association for Computational Linguistics.

Nazneen Fatema Rajani and Raymond Mooney. 2016. Combining supervised and unsupervised enembles for knowledge base population. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1943–1948, Austin, Texas, November. 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 (Volume 1: Long Papers),* Portland, Oreg., 19–24 June 2011, pages 1375–1384.

Avirup Sil and Alexander Yates. 2013. Re-ranking for joint named-entity recognition and linking. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2369–2374, Orlando, Florida, USA. ACM.

Avirup Sil, Giorgiana Dinu, and Radu Florian. 2015. The IBM systems for trilingual entity discovery and linking at TAC 2015. In *Proceedings of the Eighth Text Analysis Conference*, Gaithersburg, MD, USA. National Institute of Standards and Technology.

Valentin I Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation,* Istanbul, Turkey, 21–27 May 2012, pages 3168–3175.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden, July. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 250–259, Berlin, Germany, August. Association for Computational Linguistics.