Collective Entity Resolution with Multi-Focal Attention

Amir Globerson[*] and **Nevena Lazic** and **Soumen Chakrabarti**[†] and
Amarnag Subramanya and **Michael Ringgaard** and **Fernando Pereira**
Google, Mountain View CA, USA
gamir@post.tau.ac.il, nevena@google.com, soumen@cse.iitb.ac.in,
{asubram, ringgaard, pereira}@google.com

Abstract

Entity resolution is the task of linking each mention of an entity in text to the corresponding record in a knowledge base (KB). Coherence models for entity resolution encourage all referring expressions in a document to resolve to entities that are related in the KB. We explore attention-like mechanisms for coherence, where the evidence for each candidate is based on a small set of strong relations, rather than relations to all other entities in the document. The rationale is that document-wide support may simply not exist for non-salient entities, or entities not densely connected in the KB. Our proposed system outperforms state-of-the-art systems on the CoNLL 2003, TAC KBP 2010, 2011 and 2012 tasks.

1 Introduction

Entity resolution (ER) is the task of mapping mentions of entities in text to corresponding records in a knowledge base (KB) (Bunescu and Pasca, 2006; Cucerzan, 2007; Kulkarni et al., 2009; Dredze et al., 2010; Hoffart et al., 2011; Hachey et al., 2013). ER is a challenging problem because mentions are often ambiguous on their own, and can only be resolved given appropriate context. For example, the mention `Beirut` may refer to the capital of Lebanon, the band from New Mexico, or a drinking game (Figure 1). Names may also refer to entities that are not in the KB, a problem known as NIL *detection*.

Most ER systems consist of a *mention model*, a *context model*, and a *coherence model* (Milne and Witten, 2008; Cucerzan, 2007; Ratinov et al., 2011; Hoffart et al., 2011; Hachey et al., 2013). The mention model associates each entity with its possible textual representations (also known as aliases or surface forms). The context model helps resolve an ambiguous mention using textual features extracted from the surrounding context. The coherence model, the focus of this work, encourages all mentions to resolve to entities that are related to each other. Relations may be established via the KB, Web links, embeddings, or other resources.

Coherence models often define an objective function that includes local and pairwise candidate scores, where the pairwise scores correspond to some notion of coherence or relation strength.[1] Support for a candidate is typically aggregated over relations to all other entities in the document. One problem with this approach is that it may dilute evidence for entities that are not salient in the document, or not well-connected in the KB. Our work aims to address this issue.

We introduce a novel coherence model with an attention mechanism, where the score for each candidate only depends on a small subset of mentions. Attention has recently been used with considerable empirical success in tasks such as translation (Bahdanau et al., 2014) and image caption generation (Xu et al., 2015). We argue that attention is also desirable for collective ER due to the discussed imbalance in the number of relations for different entities.

Attention models typically have a single focus, implemented using the softmax function. Our model allows each candidate to focus on multiple mentions, and, to implement it, we introduce a novel smooth version of the multi-focus attention

[*] Currently at Tel Aviv University
[†] Currently at IIT Bombay

[1] An exception to this framework are topic models in which a topic may generate both entities and words, e.g., (Kataria et al., 2011; Han and Sun, 2012; Houlsby and Ciaramita, 2014).

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 621–631,
Berlin, Germany, August 7-12, 2016.

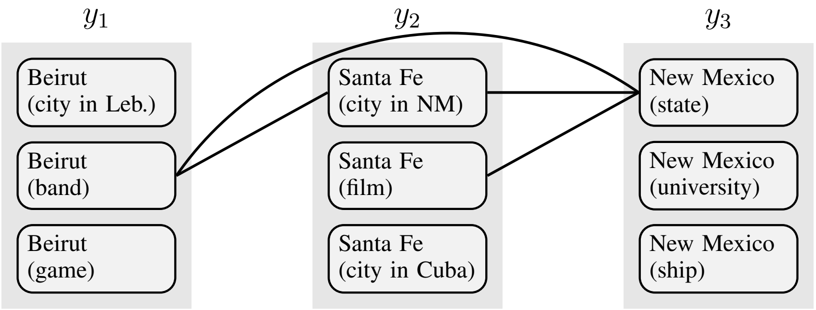

Figure 1: Illustration of the ER problem for three mentions "Beirut", "New Mexico" and "Santa Fe". each mention has three possible disambiguations. Edges link disambiguations that have Wikipedia links between their respective pages.

function, which generalizes soft-max.

Our system uses mention and context models similar to those of Lazic et al. (2015), along with our novel multi-focal attention model to enforce coherence, leading to significant performance improvements on CoNLL 2003 (Hoffart et al., 2011) and TAC KBP 2010–2012 tasks (Ji et al., 2010; Ji et al., 2011; Mayfield et al., 2012). In particular, we achieve a 20% relative reduction in error from Chisholm and Hachey (2015) on CoNLL, and a 22% error reduction from Cucerzan (2012) on TAC 2012. Our contributions thus consist of defining a novel multi-focal attention model and applying it successfully to an entity resolution system.

2 Definitions and notation

We are given a document with n mentions, where each mention i has a set of n_i candidate entities $\mathcal{C}_i = \{c_{i,1}, ..., c_{i,n_i}\}$. The goal is to assign a label $y_i \in \mathcal{C}_i$ to each mention.

Similarly to previous work, our approach to disambiguation relies on local and pairwise candidate scores, which we denote by $s_i(y_i)$ and $s_{ij}(y_i, y_j)$ respectively. The local score is based only on local evidence, such as the mention phrase and textual features, while the pairwise score is based on the relatedness of the two candidates. In Sections 3.2 and 3.3 we discuss how these scores may be parameterized and learned. Many systems (Cucerzan, 2007; Milne and Witten, 2008; Kulkarni et al., 2009) simply hardwire pairwise scores.

Coherence models typically attempt to maximize a *global* objective function that assigns a score to each complete labeling $\mathbf{y} = (y_1, \dots, y_n)$. An example of such a function is the sum of all singleton and pairwise scores for each label:[2]

$$g(\mathbf{y}) = \sum_i s_i(y_i) + \sum_i \sum_{j:j\neq i} s_{ij}(y_i, y_j). \quad (1)$$

One disadvantage of this approach is that maximizing g corresponds to finding the MAP assignment of a general pairwise Markov random field, and is hence NP hard for the general case (Wainwright and Jordan, 2008). Another limitation is that non-salient entities may be related to very few other entities mentioned in the document, and summing over all mentions may dilute the evidence for such entities. In this paper we explore alternative objectives, relying on attention and tractable inference.

3 Attention model

We now describe our multi-focal attention model. We first introduce the inference approach and optimization objective, and then provide details on how scores are calculated and learned.

3.1 Inference

As noted earlier, the global score function in Eq. (1) is hard to maximize. Here we simplify inference by decomposing the task over mentions, which makes it easy to integrate attention in terms of both inference and learning.

3.1.1 Star model

We start by considering a simple attention-free model in which inference is tractable, which we call a *star model*. For a particular mention i, the star model is a graphical model that contains y_i,

[2] The scores usually depend not only on the labels, but also on the input text. We omit this dependence for brevity.

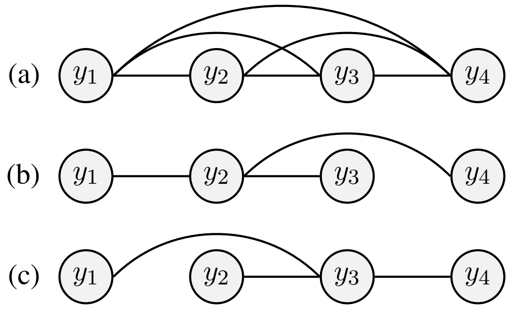

Figure 2: (a) The complete graph corresponding to Eq. (1). (b) A star shaped subgraph corresponding to y_2. This will be used to obtaining the label y_2. (c) The star graph for y_3.

all interactions between y_i and other labels, and no other interactions, as illustrated in Fig. 2.

While the star graph centered at i contains up to n variables, we will only use it to infer the label of mention i. Let $q_{ij}(y_i)$ be the support for label y_i from mention j, defined as follows:

$$q_{ij}(y_i) = \max_{y_j} s_{ij}(y_i, y_j) + s_j(y_j), \qquad (2)$$

and we also define $q_{ii}(y_i) = -\infty$ to simplify notation for later. We define the following score function for mention i:

$$f_i(y_i) = s_i(y_i) + \sum_{j:j\neq i} q_{ij}(y_i) \qquad (3)$$

and predict the label $y_i = \arg\max_y f_i(y)$.

Due to the structure of the star graph, inference is easy and can be done in $O(nC^2)$, where C is the maximum number of candidates. A similar decomposition has previously been used in the context of approximate learning for structured prediction (Sontag et al., 2011). Note that we do not view this approach as an approximation to the global problem, but rather as our inference procedure.

3.1.2 Adding attention

The score function in Eq. (3) aggregates pairwise scores for each label y_i over all mentions. In this section, we restrict this to only consider K mentions with the strongest relations to y_i.[3] Let $\text{amx}_K(\mathbf{z})$ be the sum of the largest K values in the vector $\mathbf{z} = (z_1, \ldots, z_n)$. For each label y_i, we redefine the score function to be

$$f_i(y_i) = s_i(y_i) + \text{amx}_K(\mathbf{q}_i(y_i)), \qquad (4)$$

where $\mathbf{q}_i(y_i) = (q_{i1}(y_i), \ldots, q_{in}(y_i))$ and $q_{ij}(y_i)$ is as defined in Eq. (2). The inference rule is again $y_i = \arg\max_y f_i(y)$, and the computational cost is $O(nC^2 + n\log n)$ since sorting is required.[4]

3.1.3 Soft attention

Previous work on attention has shown that it is advantageous to use a soft form of attention, where the level of attention is not zero or one, but can rather take intermediate values. Existing attention models focus on a single object, such as a single word (Bahdanau et al., 2014) or a single image window (Xu et al., 2015). In such models, it is natural to change the max function in the attention operator to a soft-max. In our case, the attention beam contains K elements, and we require a different notion of a soft-max, which we develop below.

To obtain a soft version of the function $\text{amx}_K(\mathbf{z})$, we first use an alternative definition. Denote by $\mathcal{S}$ the set $\mathbf{u} = (u_1, \ldots, u_n)$ such that $0 \leq u_i \leq 1$ and $\sum_i u_i = K$. Then $\text{amx}_K(\mathbf{z})$ is equivalent to the optimization problem:

$$\max_{\cdot \mathbf{u}\in\mathcal{S}} \mathbf{z}\cdot\mathbf{u} \qquad (5)$$

The optimization problem above is a linear program, whose solution is the sum of top K elements of $\mathbf{z}$ as required. This follows since the optimal u_i can easily be shown to attain only integral values.

Given this optimization view of $\text{amx}_K(\mathbf{z})$ it is natural to smooth it (Nesterov, 2005) by adding a non-linearity to the optimization. Since the variables are non-negative, one possible choice is an entropy-like regularizer. We shall see that this choice results in a closed form solution, and also recovers the standard soft-max case for $K = 1$. Consider the optimization problem:

$$\text{smx}_K(\mathbf{z}) = \max_{\mathbf{u}\in\mathcal{S}} \sum_i z_i u_i - \beta^{-1} \sum_i u_i \log u_i, \qquad (6)$$

where β is a tuned hyperparameter.[5] The following proposition provides a closed form solution for smx_K, as well as its gradient.

Proposition 3.1. *Assume w.l.o.g. that $\mathbf{z}$ is sorted such that $z_1 \geq \ldots \geq z_n$. Denote by R the maximum index $r \in \{1, \ldots, K-1\}$ such that:*

$$z_r \geq \beta^{-1} \log \frac{\sum_{j=r+1}^{n} \exp(\beta z_j)}{K - r} \qquad (7)$$

[3] It is possible to relax this to allow *up to* K relations, but we focus on exactly K for simplicity.

[4] Note that if $K < \log n$, we spend only nK instead of $n \log n$ time.

[5] Note that $-\sum_i u_i \log u_i$ is different from the entropy function since variables u_i sum to K and not to 1.

If this doesn't hold for any r, set $R = 0$. Then:

$$smx_K(\mathbf{z}) = \sum_{j=1}^{R} z_j + \frac{K-R}{\beta} \log \frac{\sum_{j=R+1}^{n} \exp(\beta z_j)}{K-R} \tag{8}$$

The function $smx_K(\mathbf{z})$ is differentiable with a gradient $\boldsymbol{v}$ given by:

$$v_i = \left\{ \begin{array}{ll} 1 & 1 \le i \le R \\ (K-R)\frac{\exp(\beta z_i)}{\sum_{j=R+1}^{n}\exp(\beta z_j)} & R < i \le n \end{array} \right\} \tag{9}$$

Proof is provided in the appendix.

As noted, $K = 1$ recovers the standard softmax function.[6] As $\beta \to \infty$, smx_K will approach the sum of the top K elements as expected. For finite β we have a soft version of amx_K.

Our soft attention based model will therefore consider the soft-variant of Eq. (4):

$$f_i(y_i) = s_i(y_i) + \text{smx}_K(\mathbf{q}_i(y_i)) \;, \tag{10}$$

and maximize $f(y_i)$ to obtain the label.

3.2 Score parameterization

Thus far we assumed the singleton and pairwise scores were given. We next discuss how to parameterize and learn these scores. As in other structured prediction work, we will assume that the scores are functions of the features of the input x and labels. Specifically, denote a set of singleton features for mention i and label y_i by $\boldsymbol{\phi}_i^s(x, y_i) \in \mathbb{R}^{n_s}$ and a set of pairwise features for mentions i and j and their labels by $\boldsymbol{\phi}_{ij}^p(x, y_i, y_j) \in \mathbb{R}^{n_p}$. Then the model has two sets of weights $\boldsymbol{w}_s$ and $\boldsymbol{w}_p$ and the scores are obtained as a linear combination of the features. Namely:[7]

$$\begin{aligned} s_i(y_i; \boldsymbol{w}_s) &= \boldsymbol{w}_s \cdot \boldsymbol{\phi}_i^s(x, y_i) \\ s_{ij}(y_i, y_j; \boldsymbol{w}_p) &= \boldsymbol{w}_p \cdot \boldsymbol{\phi}_{ij}^p(x, y_i, y_j) \;, \end{aligned}$$

where we have explicitly denoted the dependence of the scores on the weight vectors. See Sec. 6.2.2 for details on how the features are chosen. It is of course possible to consider non-linear alternatives for the score function, as in recent deep learning parsing models (Chen and Manning, 2014; Weiss et al., 2015), but we focus on the linear case for simplicity.

3.3 Parameter learning

The parameters $\boldsymbol{w}_s, \boldsymbol{w}_p$ are learned from labeled data, as explained next. Since inference decomposes over mentions, we use a simple hinge loss for each mention. Denote by y_i^* the ground truth label for mention i, and let $\mathbf{s}_i(y_i) \equiv (s_{i1}(y_i), \ldots, s_{in}(y_i))$. Then the hinge loss for mention i is:

$$\begin{aligned} L_i \;=\; & \max_{y_i}[s_i(y_i) + \text{smx}_K(\mathbf{s}_i(y_i)) \\ & -s_i(y_i^*) - \text{smx}_K(\mathbf{s}_i(y_i^*)) + \Delta(y_i, y_i^*)] \end{aligned}$$

where $\Delta(y_i, y_i^*)$ is zero if $y_i = y_i^*$ and one otherwise. If there are unlabeled mentions in the training data, we add those to the star graph, and maximize over the unknown labels in the positive and negative part of the hinge loss. The overall loss is simply the sum of losses for all the mentions, plus ℓ_2 regularization over $\boldsymbol{w}_s, \boldsymbol{w}_p$. We minimize the loss using AdaGrad (Duchi et al., 2011) with learning rate $\eta = 0.1$.

4 Single-link model

To motivate our modeling choices of using multi-focal attention and decomposed inference, we additionally consider a simple baseline model with single-focus attention and global inference. In this approach, which we name *single-link*, each mention i attends to exactly one other mention that maximizes the pairwise relation score. The corresponding objective can be written as

$$g^{SL}(\mathbf{y}) = \sum_i \left(s_i(y_i) + \max_j s_{ij}(y_i, y_j) \right) \tag{11}$$

where $s_{ij}(y_i, y_j) = -\infty$ if there is no relation between y_i and y_j, and we set $s_{ii}(y_i, y_i) = 0$.

While exact inference in this model remains intractable, we can find approximate solutions using max-sum belief propagation (Kschischang et al., 2001). As a reminder, max-sum is an iterative algorithm for MAP inference which can be described in terms of messages sent from model factors $g_a(\mathbf{y}_a)$ to each of their variables $y \in \mathbf{y}_a$. At convergence, each variable is assigned to the value that maximizes *belief* $b(y)$, defined as the sum of incoming messages. The message updates

[6] When we refer to the *soft-max function*, we mean the function $\beta^{-1} \log \sum \exp(\beta a_i)$, which is an often used differentiable convex upper bound of the max function (e.g., see (Gimpel and Smith, 2010)). Soft-max sometimes also refers to the activation function $\frac{\exp(a_i)}{\sum_j \exp(a_j)}$. The latter is in fact the gradient of the former (for $\beta = 1$).

[7] We again omit the dependence of the scores on the input x for brevity.

have the following form:

$$\mu_{g_a \to Y}(y) = \max_{\mathbf{y}_a \setminus y} \left[g_a(\mathbf{y}_a) + \sum_{j \neq i} q_j^{\setminus a}(y_j) \right] \quad (12)$$

where $q_j^{\setminus a}(y_j)$ is the sum of all messages to y_j except the one from factor g_a. While the single-link model contains high-order factors over n variables, computing the messages from these factors is tractable and requires sorting.

5 Related work

Ji (2016) and Ling et al. (2015) provide summaries of recent ER research. Here we review work related to the three main facets of our approach.

5.1 Coherence scores

Several systems (Milne and Witten, 2008; Kulkarni et al., 2009; Hoffart et al., 2011) use the "Milne and Witten" measure for relatedness between a pair of entities, which is based on the number of Wikipedia articles citing each entity page, and the number of articles citing both; Cucerzan (2007) has also relied on the Wikipedia category structure. Internal links from one entity page to another in Wikipedia also provide direct evidence of relatedness between them. Another (possibly more noisy) source of information are Web pages containing links (Singh et al., 2012) to Wikipedia pages of both entities. Such links have been used in several recent systems (Cheng and Roth, 2013; Chisholm and Hachey, 2015). Yamada et al. (2016) train embedding vectors for entities, and use them to define similarities.

5.2 Collective inference for ER

Optimizing most global coherence objectives is intractable. Milne and Witten (2008) and Ferragina and Scaiella (2010) decompose the problem over mentions and select the candidate that maximizes their relatedness score, which includes relations to all other mentions. Hoffart et al. (2011) use an iterative heuristic to remove unpromising mention-entity edges. Cucerzan (2007) creates a relation vector for each candidate, and disambiguates each entity to the candidate whose vector is most similar to the aggregate (which includes both correct and incorrect labels). Cheng and Roth (2013) use an integer linear program solver and Kulkarni et al. (2009) use a convex relaxation. Ratinov et al. (2011) use relation scores as features in a ranking SVM. Belief propagation without attention has been used by Ganea et al. (2015). Personalized PageRank (PPR) (Jeh and Widom, 2003) is another tractable alternative, adopted by several recent systems (Han and Sun, 2011; He et al., 2013; Alhelbawy and Gaizauskas, 2014; Pershina et al., 2015). Laplacian smoothing (Huang et al., 2014) is closely related.

5.3 Attention models

Attention models have shown great promise in several applications, including machine translation (Bahdanau et al., 2014) and image caption generation (Xu et al., 2015). We address a new application of attention, and introduce a significantly different attention mechanism, which allows each variable to focus on *multiple* objects. We develop a novel smooth version of the multi-focus attention function, which generalizes the single focus softmax-function. While some existing entity resolution systems (Jin et al., 2014; Lazic et al., 2015) may be viewed as having attention mechanisms, these are intended for single textual features and not readily extensible to structured inference.

6 Experiments

6.1 Evaluation data

CoNLL: The CoNLL dataset (Hoffart et al., 2011) contains 1393 articles with about 34K mentions, and the standard performance metric is mention-averaged accuracy. The documents are partitioned into train, test-a and test-b. Like most authors, we report performance on the 231 test-b documents with 4483 linkable mentions.

TAC KBP: The TAC KBP 2010, 2011, and 2012 evaluation datasets (Ji et al., 2010; Ji et al., 2011; Mayfield et al., 2012) include 2250, 2250, and 2226 mentions respectively, of which roughly half are linkable to the reference KB. The competition evaluation includes NIL entities; participants are required to cluster NIL mentions across documents so that all mentions of each unknown entity are assigned a unique identifier. For these datasets, we report in-KB accuracy, overall accuracy (with all NILs in one cluster), and the competition metric $B^{3+}F_1$ which evaluates NIL clustering.

6.2 Experimental setup

6.2.1 KB and entity aliases

Our KB is derived from the Wikipedia subset of Freebase (Bollacker et al., 2008), with about 4M entities. To obtain our mention prior (the probability of candidate entities given a mention), we collect alias counts from Wikipedia page titles (including redirects and disambiguation pages), Freebase aliases, and Wikipedia anchor text. 99.31% of CoNLL test-b mentions are covered by the KB, and 96.19% include the gold entity in the candidates.

We optionally use the mapping from aliases to candidate entities released by Hoffart et al. (2011), obtained by extending the "means" tables of YAGO (Hoffart et al., 2013). When released, it had 100% mention and gold recall on CoNLL, i.e. every annotated mention could be mapped to at least one entity, and the set of entities included the gold entity. However, changes in canonical Wikipedia URLs, accented characters and unicode usually result in mention losses over time, as not all URLs can be mapped to the KB (Hasibi et al., 2016, Sec. 4).

For CoNLL only, we experiment with a third alias-entity mapping derived from Hoffart et al. (2011) by Pershina et al. (2015); we call it "HP". It is not known how candidates were pruned, but it has high recall and very low ambiguity: only 12.6 on CoNLL test-b, compared to 22.34 in our KB and 65.9 in YAGO. Unsurprisingly, using only this source of aliases results in high accuracy on CoNLL (Pershina et al., 2015; Yamada et al., 2016).

Table 1 lists the statistics of the three alias-entity mappings and some of their combinations on the CoNLL test-b dataset. Table 2 provides the same statistics on the TAC KBP datasets (restricted to non-NIL mentions) for the of the YAGO+KB alias-entity mapping.

Alias map	Mention recall	Gold recall	Uniq. %	Avg. ambig.
KB	99.31	96.19	17.93	22.3
YAGO	97.17	96.30	15.50	65.9
+KB	99.84	99.51	16.28	73.6
HP	99.87	99.84	17.98	12.6
+KB	99.87	99.87	16.40	28.7
All	99.87	99.87	15.37	78.7

Table 1: Alias-entity map statistics on CoNLL test-b, 4483 gold mentions. Mention recall is the percentage of mentions with at least one known entity; gold recall is the percentage of mentions where the gold entity was included in the candidates. Unique aliases map to exactly one entity. The last column shows the number of candidates averaged over test-b mentions.

Dataset	Mention recall	Gold recall	Uniq. %	Avg. ambig.
TAC 2010	98.14	93.04	22.45	45.34
TAC 2011	98.40	89.23	27.82	49.13
TAC 2012	97.36	87.83	20.00	68.93

Table 2: YAGO+KB alias-entity map statistics on the TAC KBP datasets, restricted to non-NIL mentions.

6.2.2 Local and pairwise scores

Our baseline system is similar in design and accuracy to Plato (Lazic et al., 2015). Given the referent phrase m_i and textual context features $\mathbf{b}_i$, it computes the probability of a candidate entity as $p_i(c) \propto p(c|m_i)p(\mathbf{b}_i|c)$. The system resolves mentions independently and does not have an explicit coherence model; however, it does capture some coherence information indirectly as referent phrases are included as string context features. We experiment with several versions of the mention prior $p(c|m_i)$ as described in the previous section.

Scores for single-link model: In the single-link model, we simply set the local score for mention i and candidate c to $s_i(c) = \ln \frac{p_i(c)}{1-p_i(c)}$, so that likely candidates get positive scores. We set the pairwise score between two candidates heuristically to $s_{ij}(y_i, y_j) = \ln o(y_i, y_j) + 2.3$, where $o(y_i, y_j)$ is the number of outlinks from the Wikipedia page of y_i to the page of y_j. We consider up to three candidates for each mention for CONLL, and ten for TAC; if the baseline probability of the top candidate exceeds 0.9, we only consider the top candidate. Including more candidates did not make a difference in performance, as additional candidates had low baseline scores and were almost never chosen in practice.

Scores for attention model: Local features $\phi_i^s(x, y_i)$ for the attention model are derived from $p_i(c)$. As the attention models have no probabilistic interpretation, we inject as features $\log p_i(c)$ and $\log(1 - p_i(c))$. We set $\log 0 = 0$ by convention, and handle the case where log is undefined by introducing two additional binary indicator features for $p_i(c) = 0$ and $p_i(c) = 1$.

Edge features ϕ_{ij}^p are set based on three sources of information: (1) number of Freebase relations

System	Alias map	In-KB acc. %
Lazic (2015)	N/A	86.4
Our baseline	KB	87.9
Single link	KB	88.2
Attention	KB	**89.5**
Chisholm (2015)	YAGO	88.7
Ganea (2015)	YAGO	87.6
Our baseline	KB+YAGO	85.2
Single link	KB+YAGO	86.6
Attention	KB+YAGO	**91.0**
Our baseline	KB+HP	89.9
Single link	KB+HP	89.9
Attention	KB+HP	**91.7**
Our baseline	KB+HP*	91.9
Single link	KB+HP*	92.1
Attention	KB+HP*	92.7
Pershina (2015)	HP	91.8
Yamada (2016)	HP	**93.1**

Table 3: CoNLL test-b evaluation for recent competitive systems and our models, using different alias-entity maps. "KB+HP*" means we train and score entities using KB+HP, but output entities only in HP.

between y_i and y_j, (2) number of hyperlinks between Wikipedia pages of y_i and y_j (in either direction), and (3) number of mentions of y_i on the Wikipedia page of y_j and vice versa, after annotating Wikipedia with our baseline resolver. We cap each count to five and encode it using five binary indicator features, where the j^{th} feature is set to 1 if the count is j and 0 otherwise. Additionally, for each count c we add a feature $\log(1 + c)$. We also added a binary feature which is one if $y_i = y_j$.

We train the scores for the attention model on the 946 CoNLL train documents for CoNLL, and on the TAC 2009 evaluation and TAC 2010 training documents for TAC.

6.3 Results

CoNLL: Table 3 compares our models to recent competitive systems on CoNLL test-b in terms of mention-averaged (micro) accuracy. We also note the alias-entity map used in each system, as the corresponding gold recall is an upper bound on accuracy, and alias ambiguity determines the difficulty of the task. Therefore performance is not strictly comparable between maps.

Our baseline is slightly better than Lazic et al. (2015), but degrades after adding YAGO aliases which increase ambiguity. The attention model provides a substantial gain over the baseline, and outperforms Chisholm and Hachey (2015) by 2.3% in absolute accuracy.

The extremely low ambiguity (Tab. 1) of the HP alias mapping, coupled with guaranteed gold recall, makes the task too easy to be considered a realistic benchmark. Although we match Pershina et al. (2015) using KB+HP, for completeness, we provide the performance of our system with candidate entities restricted to those in HP (KB+HP*), but this is not equivalent to using only HP during training and inference. With KB+HP*, we outperform Pershina et al. (2015), and are competitive with recent unpublished work by Yamada et al. (2016), which uses entity and word embeddings. Including embeddings as features in our system may lead to further gains.

TAC KBP: Table 4 shows our results for the TAC KBP 2010, 2011, and 2012 evaluation datasets, where we used the KB+YAGO entity-alias map for all our experiments. To compute NIL clusters required for $B^3 + F_1$, we simply rely on the fact that our KB is larger than the TAC reference KB, similarly to previous work. We assign a unique NIL label to all mentions of an entity that is in our KB but not in TAC. For mentions that cannot be linked to our KB, we simply use the mention string as the NIL identifier. Once again, our attention models improve the performance over the baseline system in nearly all experiments, with multi-focus attention outperforming single-link. Compared to prior work, we achieve competitive performance on TAC 2010 and the best results to date on TAC 2011 and TAC 2012.

Table 5 shows two examples from the TAC 2011 dataset in which our multi-focus attention model improves over the baseline, along with the focus mentions in the document.

6.4 Effect of K and β on attention

We set the size of the multi-focus attention beam K based on accuracy on CoNLL test-a (for CoNLL) and training accuracy (for TAC). Fig. 3 shows the effect of K on the performance on CoNLL test-a dataset. Performance peaks for $K = 6$, with a sharp decrease after $K = 10$. This validates our central premise: all-pairs label coupling may hurt accuracy.

In Sec. 3.1.3 we proposed an extension of softmax smoothing to the K attention case. In our

System	In-KB acc.(%)	Overall acc.(%)	$B^{3+}F_1$
Chisholm (2015)	80.7	-	-
Ling (2015)	-	**88.8**	-
Yamada (2016)	85.2	-	-
Our baseline	84.5	87.6	83.0
Single link	84.3	87.5	82.8
Attention	**87.2**	88.7	**84.4**
Cucerzan (2011)	-	86.8	84.1
Lazic (2015)	79.3	86.5	84.0
Ling (2015)	-	-	81.6
Our baseline	81.5	86.8	84.3
Single link	82.8	87.3	84.9
Attention	**84.3**	**88.0**	**85.6**
Cucerzan (2012) R1	72.0	76.2	72.1
Cucerzan (2012) R3	71.2	76.6	73.0
Lazic (2015)	74.2	76.6	71.2
Ling (2015)	-	-	66.7
Our baseline	78.8	80.3	76.9
Single link	79.7	80.7	77.3
Attention	**82.4**	**81.9**	**78.9**

Table 4: Results on the TAC 2010 (top), TAC 2011 (middle), and TAC 2012 bottom evaluation datasets.

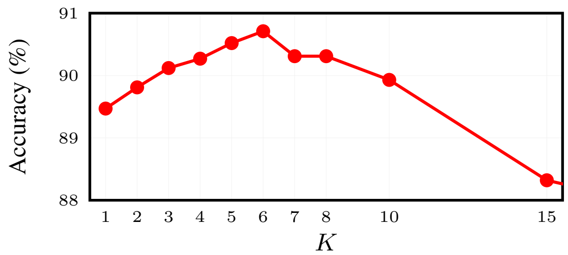

Figure 3: Effect of parameter K on entity linking accuracy. Trained on CoNLL train and tested on CoNLL test-a.

experiments we cross-validated over a wide range of β values, including $\beta = \infty$ which corresponds to taking the exact sum of K largest values. We found that the optimal value in most cases was large: $\beta = 10, 100$, or even ∞. This suggests that a *hard* attention model, where exactly K mentions are picked is adequate in the current settings.

7 Conclusion

We have described an attention-based approach to collective entity resolution, motivated by the observation that a non-salient entity in a long document may only have relations to a small subset of other entities. We explored two approaches to attention: a multi-focus attention model with tractable inference decomposed over mentions, and a single-focus model with global inference implemented using belief propagation. Our empirical results show that the methods results in significant performance gains across several benchmarks.

Experiments in varying the size of the attention beam K in the star-shaped model suggest that multi-focus attention is beneficial. It is of course possible to extend the global single-link model to the multi-focus case, by modifying the model factors and resulting messages. However, the simplicity of the star-shaped model, its empirical effectiveness, and ease of learning parameters make it an attractive approach for easily incorporating attention into existing resolution models. The model can also readily be applied to other structured prediction problems in language processing, such as selecting antecedents in coreference resolution.

Deep learning has recently been used in mutliple NLP applications, including parsing (Chen and Manning, 2014) and translation (Bahdanau et al., 2014). Learning the local and pairwise scores in our model using a deep architecture rather than a linear model would likely lead to performance improvements. The star-shaped model is particularly amenable to this architecture, as it can be implemented via a feed-forward sequence of operations (including sorting, which can be implemented with soft-max gates).

Finally, one may consider a more elaborate model in which attention depends on the current state of the system; for example, the state can summarize the mention context. The dynamics of the underlying state can be modeled by recurrent neural networks or LSTMs (Bahdanau et al., 2014).

In conclusion, we have shown that attention is an effective mechanism for improving entity resolution models, and that it can be implemented via a simple inference mechanism, where model parameters can be easily learned.

8 Proof of Proposition 3.1

Begin with the optimization problem in Eq. (6). Introduce the following Lagrange multipliers: λ for the $\sum_i u_i = K$ constraint, and $\alpha_i \geq 0$ for the $u_i \leq 1$ constraint. We can ignore the $u_i \geq 0$ constraint, as it will turn out to be satisfied. Denote

Sentence with mention	Entity	Attn. focus mentions
Caroline has dropped her name from consideration for the seat that Hillary has left vacant.	base: Caroline (given name) attn: Caroline Kennedy	Democratic Party New York Robert Kennedy
Chris Johnson had just 13 tackles last season, and the Raiders currently have have 11 defensive backs on their roster.	base: Chris Johnson (running back) attn: Chris Johnson (cornerback)	Oakland Raiders Oakland Raiders Oakland Raiders

Table 5: Examples of gains by our algorithm, showing the resolved mention, the entities it resolves to in the baseline and the attention models, and the mentions in the document that are attended to (here $K = 3$). In the first example, the baseline labels the mention "Caroline" as the given name, whereas the attention model attends to mentions that identify it as the diplomat Caroline Kennedy. In the second example, both models resolve "Chris Johnson" to football players, but the attention model finds the correct one by attending to three mentions of his former team, the Oakland Raiders.

the corresponding Lagrangian by $L(\mathbf{u}, \lambda, \alpha)$. We will show the result by using the dual $g(\lambda, \alpha) = \max_{\mathbf{u}} L(\mathbf{u}, \lambda, \alpha)$ and the fact that the solution of Eq. (6) is $\min_{\lambda,\alpha} g(\lambda, \alpha)$.

Maximizing L with respect to u_i yields:

$$u_i = e^{\beta z_i - 1 + \beta\lambda - \beta\alpha_i} \tag{13}$$

From this we can obtain the convex dual $g(\lambda, \alpha)$, and after minimizing over λ we arrive at:

$$g(\alpha) = K\beta^{-1} \log \frac{\sum_i e^{\beta z_i - \beta\alpha_i}}{K} + \sum_i \alpha_i \tag{14}$$

Next, we maximize the above with respect to $\alpha \geq 0$. Introduce Lagrange multipliers γ_i for the constraint $\alpha_i \geq 0$ and the corresponding Lagrangian $\bar{L}(\alpha, \gamma)$. We propose a solution for α, γ and show that it satisfies the KKT conditions. Minimizing $\bar{L}$ wrt α we can characterize the optimal γ as:

$$\gamma_i = -K \frac{e^{\beta z_i - \beta\alpha_i}}{\sum_i e^{\beta z_i - \beta\alpha_i}} + 1 \tag{15}$$

Set α_i as follows:

$$\alpha_i = \begin{cases} z_i - \frac{1}{\beta} \log \frac{\sum_{i=R+1}^{n} e^{\beta z_i}}{K-R} & 1 \leq i \leq R \\ 0 & R < i \leq n \end{cases} \tag{16}$$

It can now be confirmed that the α, γ from Equations 16 and 15 satisfy the KKT conditions. Plugging the α value into $g(\alpha)$ yields the solution in the proposition. Differentiability follows from Nesterov (2005) and the gradient is u_i in Eq. (13).

References

[Alhelbawy and Gaizauskas2014] Ayman Alhelbawy and Robert Gaizauskas. 2014. Graph ranking for collective named entity disambiguation. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, ACL 14, pages 75–80.

[Bahdanau et al.2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Bollacker et al.2008] Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. ACM.

[Bunescu and Pasca2006] Razvan C. Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proc. 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 06.

[Chen and Manning2014] Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

[Cheng and Roth2013] Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *EMNLP Conference*, pages 1787–1796.

[Chisholm and Hachey2015] Andrew Chisholm and Ben Hachey. 2015. Entity disambiguation with web links. *Transactions of the Association for Computational Linguistics*, 3:145–156.

[Cucerzan2007] Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proc. of EMNLP-CoNLL 2007*, pages 708–716.

[Cucerzan2012] Silviu Cucerzan. 2012. The MSR system for entity linking at TAC 2012. In *In Proc. of the Text Analysis Conference*, TAC 12.

[Dredze et al.2010] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010.

Entity disambiguation for knowledge base population. In *Proc. of the 23rd International Conference on Computational Linguistics*, COLING 10, pages 277–285.

[Duchi et al.2011] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

[Ferragina and Scaiella2010] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities). In *Proc. of the 19th ACM International Conference on Information Knowledge and Management*, CIKM 10, pages 1625–1628. ACM.

[Ganea et al.2015] Octavian-Eugen Ganea, Marina Horlescu, Aurelien Lucchi, Carsten Eickhoff, and Thomas Hofmann. 2015. Probabilistic bag-of-hyperlinks model for entity linking. *arXiv preprint arXiv:1509.02301*.

[Gimpel and Smith2010] Kevin Gimpel and Noah A Smith. 2010. Softmax-margin crfs: Training log-linear models with cost functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 733–736. Association for Computational Linguistics.

[Hachey et al.2013] Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with Wikipedia. *Artificial Intelligence*, 194(0):130 – 150.

[Han and Sun2011] Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *ACLHLT 11*. ACL.

[Han and Sun2012] Xianpei Han and Le Sun. 2012. An entity-topic model for entity linking. In *EMNLP-CoNLL*, pages 105–115.

[Hasibi et al.2016] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2016. On the reproducibility of the TagMe entity linking system. In *Advances in Information Retrieval*, pages 436–449. Springer.

[He et al.2013] Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL 13, pages 30–34.

[Hoffart et al.2011] 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 *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP11. ACL.

[Hoffart et al.2013] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

[Houlsby and Ciaramita2014] Neil Houlsby and Massimiliano Ciaramita. 2014. A scalable Gibbs sampler for probabilistic entity linking. In *Advances in Information Retrieval*, pages 335–346. Springer.

[Huang et al.2014] Hongzhao Huang, Yunbo Cao, Xiaojiang Huang, Heng Ji, and Chin-Yew Lin. 2014. Collective tweet wikification based on semi-supervised graph regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 380–390, Baltimore, Maryland, June. Association for Computational Linguistics.

[Jeh and Widom2003] Glen Jeh and Jennifer Widom. 2003. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279. ACM.

[Ji et al.2010] Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Proc. of the 3rd Text Analysis Conference*, TAC 10.

[Ji et al.2011] Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC 2011 knowledge base population track. In *Proc. of the 4th Text Analysis Conference*, TAC 11.

[Ji2016] Heng Ji. 2016. Entity discovery and linking and Wikification reading list. Online. http://nlp.cs.rpi.edu/kbp/2014/elreading.html.

[Jin et al.2014] Yuzhe Jin, Emre Kiciman, Kuansan Wang, and Ricky Loynd. 2014. Entity linking at the tail: sparse signals, unknown entities, and phrase models. In *Proc. of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 453–462, New York, NY, USA. ACM.

[Kataria et al.2011] Saurabh S. Kataria, Krishnan S. Kumar, Rajeev R. Rastogi, Prithviraj Sen, and Srinivasan H. Sengamedu. 2011. Entity disambiguation with hierarchical topic models. In *Proc. of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1037–1045. ACM.

[Kschischang et al.2001] Frank R Kschischang, Brendan J Frey, and Hans-Andrea Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

[Kulkarni et al.2009] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *Proc. of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 457–466. ACM.

[Lazic et al.2015] Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2015. Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515.

[Ling et al.2015] Xiao Ling, Sameer Singh, and Daniel S Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

[Mayfield et al.2012] James Mayfield, Javier Artiles, and Hoa Trang Dang. 2012. Overview of the TAC 2012 knowledge base population track. In *Proc. of the 5th Text Analysis Conference*, TAC 12.

[Milne and Witten2008] David N. Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *Proc. of the 17th ACM Conference on Information and Knowledge Management*, CIKM 07, pages 509–518.

[Nesterov2005] Yu Nesterov. 2005. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152.

[Pershina et al.2015] Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized Page Rank for named entity disambiguation. In *Proc. 2015 Annual Conference of the North American Chapter of the ACL*, NAACL HLT 14, pages 238–243.

[Ratinov et al.2011] Lev-Arie Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACLHLT 11, pages 1375–1384. ACL.

[Singh et al.2012] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015, University of Massachusetts, Amherst.

[Sontag et al.2011] D. Sontag, O. Meshi, T. Jaakkola, and A. Globerson. 2011. More data means less inference: A pseudo-max approach to structured learning. In R. Zemel and J. Shawe-Taylor, editors, *Advances in Neural Information Processing Systems 23*, pages 2181–2189. MIT Press, Cambridge, MA.

[Wainwright and Jordan2008] Martin J Wainwright and Michael I Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

[Weiss et al.2015] David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 323–333, Beijing, China, July. Association for Computational Linguistics.

[Xu et al.2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

[Yamada et al.2016] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. *CoRR*, abs/1601.01343.