A STUDY OF COHERENCE IN ENTITY LINKING

BY

CHASE DUNCAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

Professor Dan Roth

ABSTRACT

Entity linking (EL) is the task of mapping entities, such as persons, locations, organizations, etc., in text to a corresponding record in a knowledge base (KB) like Wikipedia or Freebase. In this paper we present, for the first time, a controlled study of one aspect of this problem called coherence. Further we show that many state-of-the-art models for EL reduce to the same basic architecture. Based on this general model we suggest that any system can theoretically benefit from using coherence although most do not. Our experimentation suggests that this is because the common approaches to measuring coherence among entities produce only weak signals. Therefore we argue that the way forward for research into coherence in EL is not by seeking new methods for performing inference but rather better methods for representing and comparing entities based off of existing structured data resources such as DBPedia and Wikidata.

To my wife, Elizabeth, who makes it all happen.

ACKNOWLEDGMENTS

Thank you to Professor Dan Roth, my advisor, who taught me how to think critically and that less sleep is the key to success.

Thank you to Bill and Gail Plater whose support and inspiration made it possible for me to seek higher education.

Thanks, Mom, for getting me started.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 A BRIEF HISTORY OF THE ENTITY LINKING PROBLEM

> Internal representations that reliably correlate with states of the world, and that participate in inferences that tend to derive true implications from true premises, may be called knowledge.
>
> — Stephen Pinker

Entity linking (EL) is the task of mapping entities in text to a corresponding record in a knowledge base (KB). Entities are unique, real world objects which have names which may or may not be unique. Some example of entities are "George Washington", "IBM", "Zimbabwe" and "Mojave Desert". Some examples of non-entities are "apple", "computer", and "railroad". The surface form of an entity is the string of characters which represents the entity in a given text, e.g. "The Big Apple", "NYC", and "New York City" are all possible surface forms for the New York City, the financial capital of the United States. We refer to a (possibly ambiguous) entity in text to be a mention.

The Internet as a tool has given humanity an unprecedented ability to store and share knowledge about everything from pop culture to history to mathematics to anything we might be curious about. Yet, if we are to use this knowledge to program computers to make inferences about text, then there must be some subroutine in that process which maps the knowledge onto the text. That subroutine, with respect to entities, is EL.

Entity linking is a problem which has been studied extensively (albeit under various names)in the past decade([1],[2],[3],[4],[5], [6], [7]). Two KBs which are commonly used as targets for entity linking are Wikipedia [1] and Freebase [2]. In this paper we assume to be linking to Wikipedia. Indeed, having done so, linking to Freebase can be achieved using a simple dictionary. As a result, it is common practice to derive entity representations from knowledge in both KBs. Theoretically, one could use the algorithms discussed in this document to link to any KB. However, it is particularly the case for coherence, that we rely heavily on the relationships between entities in the KB which are described in the KB. Robust KBs with these characteristics can be very difficult to come by.

[1] https://www.wikipedia.org/

[2] https://developers.google.com/freebase/

In entity linking we assume that the input to the system is a document, a set of entities which are contained in the document and a KB to which to ground them. In general, it need not be that the mentions in the text are identified ahead of time. The closely related task of entity discovery and linking (EDL) addresses both of these problems, discovering mentions in text before grounding them to entities in a KB. Although some of the systems in this text include the discovery step [6] [3], it is not relevant to this work. That is to say, an EL system is not concerned with determining whether the string "apple" refers to a real world entity or not. Rather, the system only seeks to determine whether or not it refers to an entity in the target KB.

Figure 1.1: Much can be understood about text by simply understanding the entities.

Trumps conduct toward Europe at the NATO and G7 summits had been so adversarial and so off-putting that he managed to compel Merkel the most iconically measured, technocratic leader in the Western world to call for Europe to chart its own course.

EL is a fundamental problem in information extraction and natural language understanding. For, if it is not possible to ascribe some conceptual semantics to the entities in a text then there is no hope of understanding the information which the text is intended to convey to the reader. Conversely, simply having unambiguous knowledge of the entities in a text can tell a reader or a computer program a great deal about the meaning of a piece of text. Consider the example in figure 1.1.

Much can be understood about the meaning in the text simply by knowing to which real-world entities each of the underlined mentions refer.

1.1.1 Key Challenges in Entity Linking

There are two key challenges in entity linking: ambiguity and variability.

Ambiguity is a one-to-many problem which occurs when the surface form of an entity in text, that is the string of characters which represents an entity in text, may refer to multiple entries in the KB. Consider the three examples in figure 1.2.

In each of these cases the surface form "Chicago" refers to 3 distinct entities in the knowledge base. Let us assume now (and for the rest of the paper) that we intend to ground mentions to Wikipedia. The challenge is to determine which of these mentions should link to the Wikipedia page *Chicago* or the page *Chicago_(musical)* or *Chicago_(band)* or maybe none of these pages.

Figure 1.2: Examples of ambiguity in the surface form *Chicago*

> Following a West End debut in 1979 which ran for 600 performances, **Chicago** was revived on Broadway in 1996, and a year later in the West End.

> **Chicago** is one of the longest-running and most successful rock groups, and one of the world's best-selling groups of all time, having sold more than 100 million records.

> **Chicago** was incorporated as a city in 1837 near a portage between the Great Lakes and the Mississippi River watershed and grew rapidly in the mid-nineteenth century.

> *Trumps conduct toward Europe at the NATO and G7 summits had been so adversarial and so off-putting that he managed to compel Merkel the most iconically measured, technocratic leader in the Western world to call for Europe to chart its own course.*

Variability is characterized by entities in the KB which are referred to by different surface forms, e.g. "Chicago" and "The Windy City" both refer to Chicago, IL.

1.2 THE EL PIPELINE

Most EL systems are constituted by a pipeline of three subroutines: candidate generation, ranking, and coherence.

Candidate generation is a process which aggregates a list of potential records in the KB to be matched to the surface form. The list of candidates for a given mention can be populated based on information that is gathered from the KB itself [3], heuristic rules [4], or using a precomputed resource like CrossWikis [8].

The ranking step then re-orders the list of candidates based on contextual information. This is done using information from the KB, e.g. how many times the surface form of a mention is linked to a Wikipedia page, as well as contextual information from the document.

Finally, coherence is used to make a final assignment based on the potential candidates for each mention. This part of the pipeline is distinct from ranking in that it largely ignores context from the document and instead utilizes information from the KB and tries to evaluate

how logical the joint assignment of entities is. The score from the ranking step is usually incorporated into this process as well.

1.3 COHERENCE IN ENTITY LINKING

This work is primarily concerned with the final step in the EL pipeline, coherence. Coherent linking in EDL is that which considers the semantic relatedness of the set of entities being linked to in the KB.

It is often the case that the ranking step is heavily biased towards popular KB entities, e.g. *Michael_Jordan* vs. *Michael_I._Jordan*. This is for good reason. [3] shows that simply linking to the most popular candidate will be correct more often than not. However, this is not true language understanding and is not useful in a practical sense. The notion of coherence attempts to counterbalance this bias by considering the total linking of all mentions in the document, or some approximation of it, and reasoning about which set of links makes the most 'sense' by some measure.

1.4 RELATED WORK

Much work has been done on the entity linking problem in the ten years. To the best of our knowledge, the first cogent formulation of the problem was put forth by [2] and the general framework that we present is heavily based on this work. This model includes both a local and global component which are both based on a vector space model. The creates a link based on maximizing the agreement between contextual context of the mention in the document and the context of a candidate title's Wikipedia page.

The global component attempts to maximize agreement between candidate titles Wikipedia categories, i.e. a granular typing that is included in a Wikipedia page's metadata, and the categories of every other candidate title being considered for the other mentions in the document. The model is a purely statistical model in the sense that there is no machine learning involved.

GLOW, for global Wikification, is the system proposed in [3]. GLOW differs from previous systems in that it uses machine learning to learn the parameters for the mention to candidate title and title to title similarity functions. Tsai and Roth [6] propose a model for cross lingual wikification based on multilingual word embeddings. Lazic et. al. [9] use EM to learn the parameters for a probabilistic Naive Bayes model. Globerson et. al. [7] build on this model with an attention mechanism for enforcing coherence. Cheng and Roth [5] are the first

to propose using relations to between entities for disambiguation. This idea is furthered by [4] which propose the Quantified Collected Validation model which is entirely based on coherence. There are many more.

1.5 WHAT LIES AHEAD

In chapter 2 we put forth a general model for entity linking with coherence. We discuss the entity linking problem in greater detail as well as potential similarity measures, document representations, and candidate record representations. In section 2.2 we do a deep dive on related work as it pertains to coherence in entity linking. In chapter 3 we lay out the results of the experiments we perform and finally in chapter 4 we discuss the results as well as make prescriptions for next steps in this field of inquiry.

CHAPTER 2: A GENERAL FRAMEWORK FOR COHERENCE IN ENTITY LINKING

There are two resources from which to derive the knowledge necessary to appropriately disambiguate the mentions in a document: the document itself and the KB. These signals are modeled separately in a *mention space* model and a *candidate space* model. The mention space model quantifies the similarity between the mentions in the document and the candidate entities using features derived from the document in relation to context taken from the KB. This model requires a minimum of three components: a representation for the context of the mention (the document), a representation for the candidate entity in the KB and a similarity function. The candidate space model quantifies the relationships between the candidate entities and is wholly derived from the KB. This is part of the problem which is referred to as *coherence*. Similarly to the first model it requires a schema for representing the candidates and a similarity measure over those representations.

The most general framework for entity linking is as follows:

$$\underset{(e_1,...,e_N)\in C(m_1)\times C(m_N)}{\operatorname{argmax}} \sum_{i=1}^{N} \phi(m_i, e_i) + \Psi(e_1, ..., e_N) \tag{2.1}$$

where e_i is a candidate entity for a mention m_i. $C(m_i)$ is the set of candidates for the mention m_i. $\phi(m_i, e_i)$ is a scoring function over the the domain of mention-entity pairs. $\Psi(e_1, ..., e_N)$ is the coherence function over the joint assignment of each of the entities to each mention in the document.

ϕ may include lexical features such as context words and typing. Ψ includes features which are derived from the KB. These may also be referred to as *local features* and *global features*, respectively [3].

Quantifying coherence over a joint assignment of entities is NP-hard [2]. Thus we typically quantify coherence using a pairwise scoring function which computes the similarity between some representation of pairs of entities and take the sum of pairwise measurements of each entity pair such as

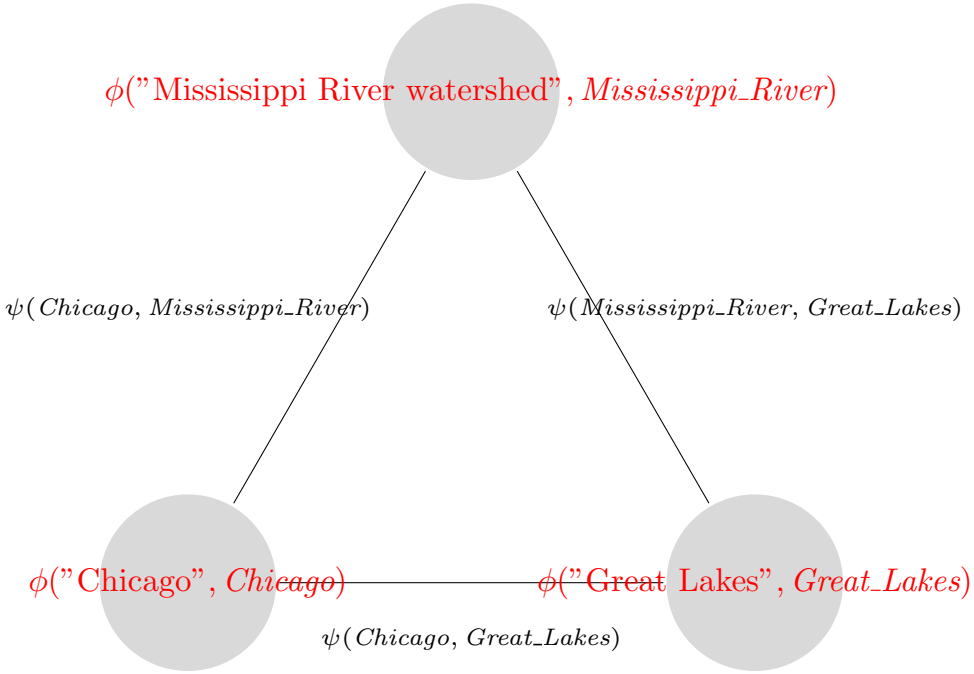
$$\underset{(e_1,...,e_N)\in C(m_1)\times C(m_N)}{\operatorname{argmax}} \sum_{i=1}^{N} \phi(m_i, e_i) + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \psi(e_i, e_j) \tag{2.2}$$

this approximation is intuitive when considering the graphical nature of the knowledge base. It is natural to think of $\phi(m_i, e_i)$ as function which assigns a weight to a node in a graph and $\psi(e_i, e_j)$ as a function which assigns a weight to and edge in a graph. Then

the problem reduces to finding a the graph which maximizes the sum of the weights on its vertices and edges. This formulation is equivalent to finding an assignment of entities from the candidate set for each mention in the document, $(e_1, ..., e_N) \in C(m_1) \times C(m_N)$, which maximizes this objective function.

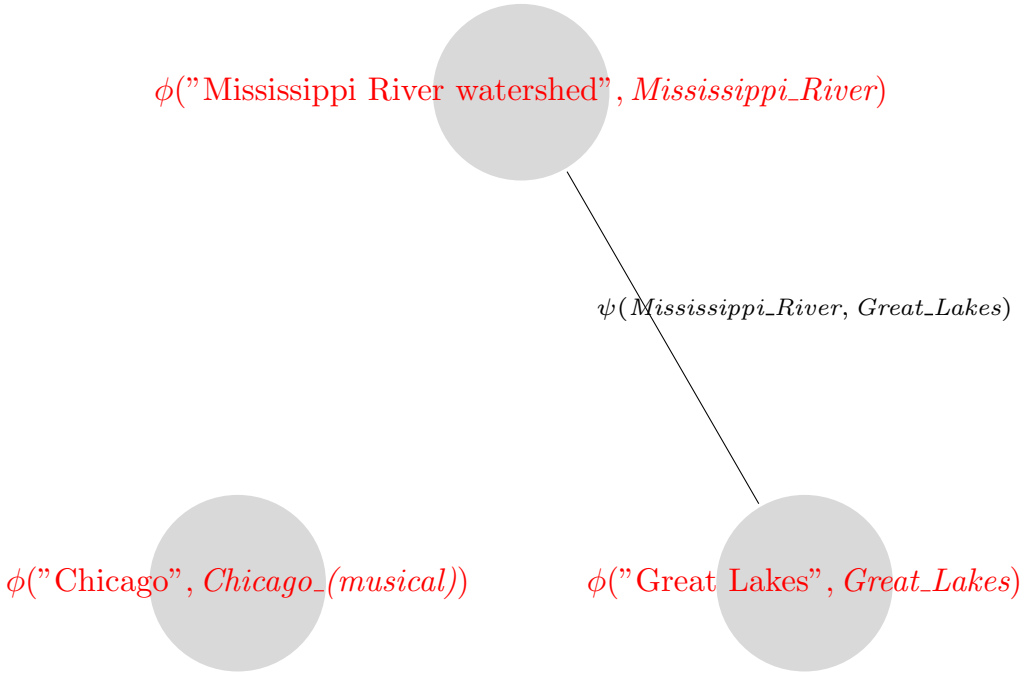
See the figures 2.1 and 2.2 for examples of this idea.

Figure 2.1: Choosing between $\Gamma = (Chicago, Great\_Lakes, Mississippi\_River)$ and $\Gamma = (Chicago\_(musical), Great\_Lakes, Mississippi\_River)$.



Cucerzan notes that even in this approximate form "the quality of an assignment of an entity to a surface form depends on all the other assignments made, which makes this a difficult optimization problem." [2] Therefore, almost all systems include one more approximation which allows a linking decision via coherence for each mention individually using an

Figure 2.2: Choosing between $\Gamma = (\textit{Chicago}, \textit{Great\_Lakes}, \textit{Mississippi\_River})$ and $\Gamma = (\textit{Chicago\_(musical)}, \textit{Great\_Lakes}, \textit{Mississippi\_River})$.



approximation for the best links of the other mentions in the document. This approximate context is called the disambiguation context by [3] and we adopt that name here.

This yields the problem statement:

$$\underset{(e_1,\ldots,e_N)\in C(m_1)\times C(m_N)}{\operatorname{argmax}} \sum_{i=1}^{N}[\phi(m_i, e_i) + \sum_{\substack{j\in\Gamma' \\ j\neq i}} \psi(e_i, e_j)] \tag{2.3}$$

This final formulation allows each mention to be evaluated separately which makes the problem much more easy to solve computationally.

In an effort to simplify the notation, let $\Gamma = (e_1, ..., e_N) \in C(m_1) \times C(m_N)$ be an assignment of titles to mentions in a document. Finally, let $\phi_{ii} = \phi(m_i, e_i)$ be the scoring function over the context in the document of mention m_i and the in-KB context of the candidate e_i and let $\psi_{ij} = \psi(e_i, e_j)$ be the scoring function over the two candidate titles e_i and e_j. The disambiguation context will be referred to as Γ'.

In this work we are primarily concerned with coherence which is defined by ψ_{ij} and Γ'. With all this we now state the final form of the general framework of entity linking with coherence:

$$\underset{\Gamma}{\operatorname{argmax}} \sum_{i=1}^{N} [\phi_{ii} + \sum_{\substack{j \in \Gamma' \\ j \neq i}} \psi_{ij}] \tag{2.4}$$

We will see that the key differences among the various EL systems are as follows:

1. The scoring function ϕ_{ii}, which includes a schema for document representation, and schema for candidate representation and one or more similarity functions which operate upon them.

2. The scoring function ψ_{ij}, which includes a schema for candidate representations and one or more similarity functions which operate on them.

3. The disambiguation context Γ', the approximation of Γ which is used in the coherence term of 2.4, i.e. ψ_{ij}.

From the above we conclude that EL task is essentially a matter of finding good representations for the document context and candidate entities vis-á-vis each other, good representations of the candidate entities vis-á-vis each other, similarity functions which operate on them and which candidate titles to choose for computing a 'coherent' assignment.

For example, consider a very simple bag of words representation for the document context and Wikipedia pages, $BOW(d)$ which yields a sparse vector representation of a document. Then we could use cosine-similarity to compute the similarity of these representations. Then we could define an EL system as follows:

$$\underset{\Gamma}{\operatorname{argmax}} \sum_{i=1}^{N} [cossim(BOW(d_{m_i}, e_i)) + \sum_{\substack{j \in \Gamma' \\ j \neq i}} cossim(BOW(e_i, e_j))] \tag{2.5}$$

where d_{m_i} is the textual context of the mention m_i, we could take it to be the whole document or words near the mention, etc. This may not be a very good EL system but it

exemplifies the basic requirements. Now we discuss more (and possibly less) sophisticated components.

2.1 COMMON MEASURES OF COHERENCE

The precise structure of ψ_{ij} varies among systems. Some systems express coherence as linear combination of multiple features, e.g. [7] and [3] which deduce weights for the features using machine learning. Other systems model coherence using statistical models built from Wikipedia and Freebase like [2], [10], [11] and [12]. In either paradigm there is a common set of signals which are exploited to measure coherence. We describe them in detail in this section.

2.1.1 Normalized Google Distance

A pervasive measure of semantic relatedness between Wikipedia pages is interchangeably referred to as normalized Google distance or the Milne and Witten measure. [12] were the first to propose measuring the similarity between Wikipedia pages using the normalized Google distance (NGD) measure [13]. NGD is a semantic similarity measure between sets of keywords based on the number of pages returned by using each keyword set as Google search, the feature of the algorithm from which the name is derived.

[12] adapt this measure to compute the semantic similarity between Wikipedia pages based on incoming links to each page. Namely,

$$NGD(t_1, t_2) = 1 - \frac{\log\Big(\max\Big(\Big|T_1\Big|, \Big|T_2\Big|\Big)\Big) - \log\Big(\Big|T_1\Big| \cap \Big|T_2\Big|\Big)}{\log\Big(\Big|W\Big|\Big) - \log\Big(\min\Big(\Big|T_1\Big|, \Big|T_2\Big|\Big)\Big)} \tag{2.6}$$

where t_1 and t_2 are Wikipedia page, T_1 and T_2 are the sets of incoming links to each page (or outgoing links), and W is the the total number of articles on Wikipedia.

As a point of clarification, a 'link' is a http hyperlink from one page to another in Wikipedia. These links constitute the edges in the knowledge graph. We refer to all of the hyperlinks which link to a page collectively as the *inlinks* and the set of hyperlinks which link from a page as the *outlinks*. Therefore, the candidate representation in this context is a set of Wikipedia pages.

Let's look closer at what this expression tells us. First, NGD is built on the assumption that shared inlinks or outlinks can be taken as a proxy for similarity. This is an application of the distributional hypothesis, "a word can be known by the company it keeps" [14]. In

this case the "word" is the title and its company are its inlinks and outlinks, an assumption that has merit because of the thriving community of contributors to Wikipedia.

Therefore, we take the distance between two Wikipedia pages to be the ratio of the number of links which the larger of the two pages, call it T_{max}, does not share with the smaller page, call it T_{min}, by the total number of pages which the larger page could potentially link to. If we analyze the edge cases we see that NGD is maximal, 1, when T_{max} links to every page in Wikipedia except those which T_{min} links to. On the other hand, it is minimal, 0, if the T_{max} and T_{min} link to the exact same set of Wikipedia pages.

2.1.2 Explicit Counts of Outlinks and Inlinks

As discussed in 2.1 outlinks are the set of Wikipedia which a given page links to and inlinks are the set of Wikipedia pages which link to a given page. Some models simply use the number of shared outlinks (or inlinks) as a feature. Namely, in [7] where it is a feature among a suite of features.

Inlink and outlink counts have also been used to generate a probability of $p(t_1|t_2)$ such as in [15] which uses this coherence measure:

$$p(t_1|t_2) = \frac{I_1 \cap I_2}{I_2} \tag{2.7}$$

where I_i are the inlinks to page t_i.

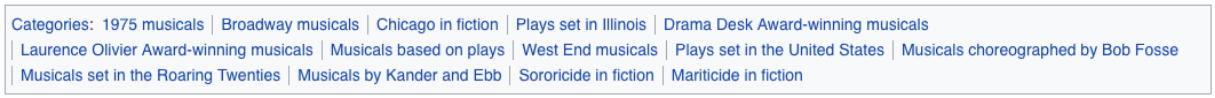
Again the representation of the candidate pages is a set of Wikipedia titles. What has changed from NGD is the similarity measure.

2.1.3 Co-occurrence of Wikipedia Titles

Two titles, t_1 and t_2, are said to co-occur in Wikipedia when there exists a Wikipedia page, t_s, for which $t_1 \in outlinks(t_s)$ and $t_2 \in outlinks(t_s)$ where $outlinks(t_s)$ is the set of all Wikipedia pages which t_s links to. There are at least a few ways to quantify co-occurrence. These may include a simple indicator function signaling that the entities do (or do not) co-occur. It may be the sum of the total number of time which two titles are linked from the same pages. Or it may be the product of the same counts.

This is a relatively new metric for coherence having only been seen in [7]. In this case, the co-occurrence statistics are taken over text from Wikipedia which has already been annotated by an EL system. This is thought to be useful since most entities are only linked the first time they are seen in an article. Therefore it's only possible to measure co-occurrence as

Figure 2.3: The categories for the Wikipedia page *Chicago_(musical)*.



a binary phenomena, i.e. two pages co-occur in a document or not. However, annotating with an EL system first allows us to measure the magnitude with which entities co-occur. Of course, this runs the risk of introducing the bias of the model. In our experiments, only one-third of the total pages in Wikipedia are linked in this way. Yet, less than 2% of the articles on Wikipedia are *orphaned*[1], i.e no other articles link to them. This suggests that this particular statistic will be heavily biased by the system which creates the annotation.

2.1.4 Wikipedia Categories

Wikipedia categories are a typing mechanism which is used to cluster Wikipedia pages which contain similar subject matter. Like all aspects of Wikipedia they are user managed tags. See 2.3 for an example of the categories for the Wikipedia page for the musical *Chicago*.

2.1.5 Pointwise Mutual Information

Pointwise Mutual Information (PMI), a well-known measure of association from information theory, has been used to measure similarity of Wikipedia titles [3]. In its general form, PMI is the ratio of the joint probability of two events by the product of each the event's prior. This is adapted for this application using inlinks:

$$PMI(t_1, t_2) = \frac{|I_1 \cap I_2|/|W|}{|I_1|/|W||I_2|/|W|} \tag{2.8}$$

Again, in this case, the candidate representation is a set of Wikipedia titles. An idea which has not been explored is using PMI in conjunction with co-occurrence statistics.

2.1.6 Relations

Cheng and Roth [5] introduced the idea of using information regarding the relationships between entities which can be harvested from the KB. Links between records in the KB is a strong indicator of similarity, at least in a topical sense, but many KBs such as Freebase, WikiData, and DBPedia include explicit information about how entities in the KB relate to

[1] https://en.wikipedia.org/wiki/Category:Orphaned_articles

each other. In general, these facts are viewed as triples of the form (e_h, r, e_t) where e_h is the head entity in the fact, e_t is the tail entity and r is the unidirectional relationship of e_h to e_t [4]. Another way to think of this tuple is as $(subject, predicate, object)$.

Figure 2.4 gives some examples of Freebase relations.

Figure 2.4: Example relation facts from Freebase.

Michelle_Rodriguez *award_winner* Naveen_Andrews

Australia_national_soccer_team *position* Midfielder

Maldives_national_football_team *position* Forward_(association_football)

Bryan_Singer *film* Star_Trek:_Nemesis

BAFTA_Award_for_Best_Original_Screenplay *nominated_for* Philadelphia_(film)

Danny_DeVito award_nominee Guy_Pearce

Harpsichord *role* Violin

Academy_Award_for_Best_Foreign_Language_Film *ceremony* 61st_Academy_Awards

Relations introduce challenges in determining how to best quantify them for the task.

2.2 AN INDEPTH LOOK AT RELATED WORK

Now we consider some of the most relevant entity linking models and how they address the problem of coherence. Further, we show how they all reduce to the general framework for coherence in entity linking.

2.2.1 Cucerzan, 2007

In [2] a sparse representation is used to represent the the context of the document and the contexts of the Wikipedia pages which are being considered as candidates for linking. To calculate the similarity between these representations, a simple vector product is used.

The context vector for a document $d = \{d_1, d_2, ..., d_M\} \in \mathbb{N}^M$ counts the occurrences of a context from a context lexicon of size M, i.e. d_i is number of occurrences of ith element in the lexicon. The contexts in the lexicon are the appositives in the page names, such as "musical" in *Chicago_(musical)*, the outlinks in the first paragraphs of each Wikipedia page, which tends to be a summary of the Wikipedia page, and all the Wikipedia pages which

jointly link with each other. Then a document is represented by it's context with a vector d such that:

$$d^i = \begin{cases} 1, & c_i \text{ is a context for d} \\ 0, & \text{otherwise} \end{cases} \tag{2.9}$$

Additionally, Wikipedia categories, discussed in 2.1.4, are used to represent candidate titles. This is once again accomplished using a sparse representation $\delta_e|_T \in \{0,1\}^N$ where $T = \{t_1, t_2, ..., t_N\}$ are Wikipedia categories. A context vector $\delta_e|_C \in \{0,1\}^M$ is also used to represent entities. Thus two vectors are used to represent entities:

$$\delta_e^i|_C = \begin{cases} 1, & c_i \text{ is a context for e} \\ 0, & \text{otherwise} \end{cases} \quad \delta_e^j|_T = \begin{cases} 1, & t_j \text{ is a category for e} \\ 0, & \text{otherwise} \end{cases} \tag{2.10}$$

Then as in 2.2 an assignment is made according to

$$\underset{\Gamma}{\text{argmax}} \sum_{i=1}^{N} \langle \delta_e^i|_C, d\rangle + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \langle \delta_e^i|_T, \delta_e^j|_T\rangle \tag{2.11}$$

in which $\langle \cdot, \cdot \rangle$ is the vector product.

This formulation suffers from the same issues of tractability which were discussed earlier. Namely, computing the best assignment among all assignments according to all pairs of candidates grows exponentially in the number of mentions. Therefore, the disambiguation context, Γ', is taken to be the sum of all possible assignments for each mention. As in,

$$\underset{\Gamma}{\text{argmax}} \sum_{i=1}^{N} \langle \delta_e^i|_C, d\rangle + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \langle \delta_e^i|_T, \sum_{e\in C(m_j)} \delta_e|_T\rangle \tag{2.12}$$

where $C(m_j)$ is the set of all candidates for the mention, m_j. If we pull out the summation over the mentions we have

$$\underset{\Gamma}{\text{argmax}} \sum_{i=1}^{N} [\langle \delta_e^i|_C, d\rangle + \sum_{\substack{j=1 \\ j\neq i}}^{N} \langle \delta_e^i|_T, \sum_{e\in C(m_j)} \delta_e|_T\rangle] \tag{2.13}$$

i.e., 2.3, as expected.

It's worth nothing that very little information is being used about the document. Only the occurrences of appositives from titles.

2.2.2 Illinois Wikifier (GLOW)

GLOW [3] uses machine learning to learn the parameters for a scoring function between a mention and a candidate title, $\phi(m_i, e_i)$, and the parameters for a scoring function between two titles, $\psi(e_i, e_j)$. The features of $\phi(m_i, e_i)$ are cosine similarities between various representations of the mention context and the candidate page context. These include TF-IDF summaries of the candidate Wikipedia page, a token window around the mention, the context in which the candidate is linked within Wikipedia and the document itself.

The features of $\psi(e_i, e_j)$, the coherence scoring function, are PMI and NGD of inlinks and outlinks. Thus the scoring function is a linear combination of the standard similarity measures over candidates which are represented as sets of Wikipedia pages.

The GLOW objective is as follows:

$$\Gamma* \approx \underset{\Gamma}{\operatorname{argmax}} \sum_{i=1}^{N} [\phi(m_i, e_i) + \sum_{e_j \in \Gamma'} \psi(e_i, e_j)] \tag{2.14}$$

There a two key differences between GLOW and Cucerzan's model [2].

The scoring functions are more complicated in GLOW. Whereas the scoring function in [2] was simply a dot product overly relatively simple representations of the KB records and the document, [3] learns a ranking over features such that each feature has some representation of candidate entities or the mention context.

Another key difference between the models is in how the disambiguation context, Γ', is aggregated. In [2], Γ' is sum of all potential candidates for a given mention. In GLOW the disambiguation context is the top candidate yielded by the ranking model ϕ_{ii}.

In this way the solution is chosen over a two stage process: first, the candidates are ranked for each mention using local signals according to ϕ_{ii}. Then Γ' is taken as the best candidate from this stage and the mentions are linked again, possibly changing the initial link, in light of this new information.

It's interesting to note that, although it is not mentioned explicitly in the paper, inspection of the source code reveals that Illinois Wikifier includes a rudimentary attention mechanism. For calculating coherence, the system only admits mentions which are within 2000 characters (approximately 200 tokens, or a paragraph) of the mention.

GLOW uses NGD and PMI over both inlinks and outlinks of candidate titles as features in the pairwise similarity measure, ψ_{ij}. In a similar way to [2], GLOW only considers coherence for titles which are mutually linked. Conveniently, the GLOW objective 2.14 has an equivalent form to the general framework 2.3.

2.2.3 Vinculum

Vinculum [11] has a coherence model which is similar to the global component of GLOW but with two key simplifications. The inference step is the same, i.e. a greedy assignment of candidates which maximize the sum of ϕ_{ii} and ψ_{ij}. However, Γ' is simply the best candidates for each mention as given by the candidate generation step. That is to say, ϕ_{ii} has no bearing on Γ' in Vinculum. Vinculum uses CrossWikis for candidate generation.

Additionally, the ψ_{ij} is either NGD, a relational similarity measure, or the average of both but, in any case, there is no learning involved.

The Vinculum coherence model is

$$\frac{1}{|P_d| - 1} \sum_{p \in P_d \setminus p_m} \phi(p, c), c \in C_m \tag{2.15}$$

where $P_d \setminus p_m$ is the set of intermediate best-links for each of the other mentions in the document. Since P_d is created dynamically for each mention, we evaluate different strategies for iterating through the mentions based on confidence in the score from ϕ_{ii}.

As previously mentioned, NGD and relations are used as features for ψ_{ij}. NGD is is implemented as in 2.6. The relational ψ_{ij} is defined as

$$\psi_{REL}(e_i, e_j) = \begin{cases} 1, & \exists r \text{ s.t. } (e_i, r, e_j) \text{ or } (e_i, r, e_j) \in F \\ 0, & \text{otherwise} \end{cases} \tag{2.16}$$

where F is Freebase and (e_i, r, e_j) are relational facts as described in section 2.1.6. Therefore, this is simply the proportion of the number of other candidates with which a given candidate has with the other candidates in Γ'.

They show their best performing ψ_{ij} to be the average of NGD and ψ_{REL}.

2.2.4 Globerson

Globerson et. al. [7] introduce the concept of attention [16]to the coherence model. Heretofore, the process for pruning Γ' was entirely a function of the candidates. The big idea behind attention is that it is also possible to intelligently choose the disambiguation context by only using mentions whose candidates are most relevant in the coherence scoring process.

The coherence model (ψ_{ij}) is built using Lazic et. al. [9] for ϕ_{ii}. The system considers one mention at a time for linking, i.e. there is no joint linking. The primary thrust of the work is a 'soft' attention model which dynamically adjusts the size of the attention window based

on the coherence scores of the other mentions in the document with respect to the mention being evaluated for linking.

However, in the results [7] says that a hard attention model, i.e. a fixed window size for attention, performs best in most cases.

The Globerson model suffers from some fairly considerable drawbacks. First, it is a supervised model which requires training data beyond Wikipedia. In practice this requirement makes the model quite expensive relative to other models which perform similarly well. Second, the model is fairly complicated and requires a lot of effort to implement. These issues will cause the model to be impractical in many EL environments.

CHAPTER 3: EXPERIMENTS AND ANALYSIS

The baseline model which we use to evaluate the impact of coherence in entity linking is the model proposed by (Tsai and Roth) [6]. This model uses word embeddings to represent document context as well as Wikipedia pages. It uses cosine similarity between these representations and learns and SVM ranking model which learns a linear combination of three similarity functions. The first computes the similarity between the dense representation of the candidate and a dense representation of the document based on other mentions in the document. Another computes the similarity between the dense representation of the candidate and a dense representation of the document based on lexical features. And finally one similarity function which is parameterized by the candidate vector and a vector which is the sum of all previously linked candidates. Note that the last component has some elements of coherence.

We evaluate on two datasets, TAC-KBP 2016 [1] and TAC-KBP 2017 [2]. These tasks include linked nominals and out-of-KB links, called nil-linking. We only evaluate on in-KB (non-nil), named entities. Linking nominals is mostly a co-reference task and downstream from coherence. Nil-linking is very important, and it's possible that coherence maybe used to improve the results on this task, but, again, it is downstream from coherence so we do not evaluate on that subset of the data.

This leaves 5,378 linkable mentions in the TAC-KBP 2016 dataset and 3,569 linkable mentions in the TAC-KBP 2017 dataset.

3.1 SIMILARITY MEASURES AND REPRESENTATIONS

3.1.1 Baseline and Recall

Table 3.1 has the results for the baseline model on the datasets, TAC-KBP 2016 and TAC-KBP 2017. We define recall@k as being the proportion of times that the ranking model has the correct candidate in the first k candidates. In effect recall@MAX is the upper bound on how well coherence can perform where MAX is the maximum number of candidates in any list. This model restricts candidate lists to the size of 20. Therefore, recall@20 is the aspirational score. The game of coherence is to get the score in last column of table 3.1 to be as close as possible to score in the first column.

[1] https://tac.nist.gov//2016/KBP/
[2] https://tac.nist.gov//2017/KBP/

Recall@5 and recall@2 are useful data points for understanding how we can assemble Γ'. For instance, [5] operates under the assumption that we can take assume the ranker is good enough that we only need to consider the top 2 candidates. However, we can see in our experiments that we may greatly hinder how well a model can perform based on this assumption.

Table 3.1: The performance of the baseline model on the TAC-KBP 2016 and TAC-KBP 2017 datasets. We measure the recall at top 20, top 5, and top 2 candidates as well as the precision, i.e. the accuracy at 1.

	@20	@5	@2	@1 (Precision)
TAC-KBP 2016	0.910	0.884	0.867	0.861
TAC-KBP 2017	0.880	0.847	0.822	0.822

3.1.2 NGD, PMI and REL

In our experiments we calculate NGD just as in 2.6. PMI is calculated as in 2.8. Additionally we apply the standard logistic function to the value returned by 2.8 in order to make a more sensible combination with it and the other scores. Therefore, PMI in our experiments is

$$PMI = \frac{1}{1 + e^{-pmi(t_1,t_2)}} \tag{3.1}$$

where

$$pmi(t_1, t_2) = \frac{|I_1 \cap I_2|/|W|}{|I_1|/|W||I_2|/|W|} \tag{3.2}$$

For both PMI and NGD the candidates are represented as sets of outlinks.

We use the Vinculum formulation for relational similarity (see 2.16). The representation of the candidates are sparse vector representations over a vocabulary of relations. For the relation counts themselves we use the FB15K-237 dataset [17] from which we use the training, test, and validation sets. This dataset represents a pruned version of the complete set of Freebase relations which removes redundant data. It's worth mentioning that there only 14407 unique titles in the data which means only a fraction of the total number of Wikipedia pages occurs in a relation tuple.

Vinculum found average of NGD and relations to perform best. We evaluate this combination as well an average of all three measures.

The results of all of these experiments is in table 3.2.

Table 3.2: A comparison of the various coherence measures. BASE is the baseline model, NGD is Normalized Google Distance, PMI is Pointwise Mutual Informations, and REL is the relations measure.

Similarity Measure	Representation Schema	TAC-KBP 2016	TAC-KBP 2017
BASE	N/A	0.867	0.822
NGD	Outlinks	**0.871**	0.823
PMI	Outlinks	0.861	0.823
REL	Freebase Relation	0.861	**0.846**
NGD+REL	Outlinks and Freebase Relations	0.861	0.823
ALL	Outlinks and Freebase Relations	0.870	0.823

Note that these scores are without the "confidence rule" discussed in section .

3.1.3 Order of Inference

When doing greedy inference in a dynamic disambiguation context, i.e. the Γ' potentially changes at each step based on greedy linking, then it's possible the order of the inference could affect the outcome. We perform tests to measure the impact of this phenomena. In our tests we find that the order of inference has little impact on the result. We evaluate a model which performs inference based on how confident the top link is from the ranker where confidence is taken to be the difference in scores between the top two candidates. The intuition is that a confident linking will be scored much higher than the next candidate in the list.

The model then performs the greedy inference in three ways: 1) in ascending order from least confident to most confident, 2) in descending order from most confident to least confident, and 3) in the order in which the mentions appear in the document, i.e. by ignoring the confidence score. The numbers included in table 3.3 use the NGD similarity measure over outlinks but the results were similar for all measures.

3.1.4 Disambiguation contexts

We investigate two mechanisms for aggregating the disambiguation context, Γ'. The simplest approximation is to take the set of all the best candidates according to the ranker and remove the candidate under consideration. This strategy is used by [3], [18] and [5] and we call it BEST. The second strategy is to use the set of *all* other candidate titles except that which is being evaluated. This is used in [7] and [2] and we call it ALL. The results of these experiments can found in table 3.4.

Table 3.3: The comparison between strategies for ordering the greedy inference. We measure confidence as being the delta between the scores the ranker assigns to the first and the second candidates in the candidate list for a mention. Then there are 3 strategies for greedy inference: evaluate the mentions from least to most confident, evaluate the mentions from most to least confident, or simply ignore the confidence and evaluate the mentions based on the order they appear in the text. In each case we dynamically reassign the top candidate for the mention being evaluated based on coherence at each step of the inference procedure. NGD is used as the coherence measure.

	High to Low Confidence	Low to High Confidence	In Order
TAC-KBP 2016	0.861	0.861	0.861
TAC-KBP 2017	0.822	0.822	0.822

Table 3.4: The comparison between disambiguation context aggregation strategies. BEST creates the disambiguation context from the top candidate as given by the ranker. ALL creates the disambiguation context from all candidates.

	BEST	ALL
TAC-KBP 2016	0.861	**0.871**
TAC-KBP 2017	**0.846**	0.826
Average	**0.854**	0.848

3.1.5 The Challenge of Combining the Scoring Functions

One of the major challenges in building an EL system which incorporates coherence is figuring out how to combine the local scoring function ϕ_{ii} and the global scoring function ψ_{ij}. Through observing instances where the coherence system incorrectly changed the ranker links, we added a rule whereby the coherence system will only reassign if the value of $\phi_{ii}+\psi_{ij}$ is greater than the confidence (see 3.1.3) of the original link.

This rule significantly improved the performance of the system on the TAC data. As seen in table 3.5. This heuristic suggests that there is more to be understood about how to combine the two scoring mechanisms.

Table 3.5: The comparison between systems with and without confidence rule. In each case the relation representation is used. The confidence rule is simply that the coherence system will not change a link from the ranker unless the total score by the coherence system is greater than the confidence of the ranker.

	No Confidence Rule	Confidence Rule
TAC-KBP 2016	0.861	**0.892**
TAC-KBP 2017	0.846	**0.863**
Average	0.853	**0.877**

3.2 ATTENTION

3.2.1 Attention Experiments

We investigate attention according to various sized 'windows' of attention, i.e. various values of k. The results are found in table 3.6.

Table 3.6: The comparison between varying values of k for the size of the attention window. Each of these experiments use a relational scoring mechanism and dynamic BEST disambiguation context.

k	TAC-KBP 2016	TAC-KBP 2017
3	0.861	0.846
6	0.861	0.846
10	0.861	0.846

3.2.2 Globerson Experiments

For this work we built a version of the system outlined by Globerson et. al. in [7]. This model uses outlinks and co-occurrence counts to represent the candidates. At the most basic level it uses hinge loss to learn parameters for a linear combination of these features.

The Globerson model is parameterized by two values, k and β. Intuitively, k is the size of the attention window and β is a slackness factor which determines how 'soft' the attention is. When β is large and the attention is 'hard' then exactly k other mentions are used as support for measuring the coherence of a candidate. When the attention is soft then the attention is distribute across n other mentions such that there are values $\alpha_1 + ... + \alpha_n = k$ and the coherence window is taken as $\alpha_1 m_1 + ... + \alpha_n m_n$.

Using the parameters $k = 6$ and $\beta = 1$.and a learning rate of 0.01. For these experiments we only evaluate on mentions where the baseline model generates the correct candidate in the top 20 candidates. This simplifies the training and testing procedures. Table 3.7 shows the results from these experiments.

3.3 COHERENCE EXAMPLES

In this section we present some specific examples which illustrate how coherence in entity linking can improve the output of the ranking model.

Table 3.7: Comparing our implementation of the Globerson model with the baseline ranker. The first row shows the result from evaluating the trained model on the training data, TAC-KBP 2016. The second row shows the result of training on TAC-KBP 2016 and testing on TAC-KBP2017. In all cases we use the relational coherence scoring.

Train	Test	Baseline	Globerson Model
TAC-KBP 2016	TAC-KBP 2016	**0.945**	0.915
TAC-KBP 2016	TAC-KBP 2017	**0.917**	0.816

Figure 3.1: Example of coherence in EL on forum post.

Headline: ***Tom Merritt*** Fired from **TWiT**

"MacNut":
So it looks like **Leo's** ego got in the way again as he fired ***Tom*** today.

"Jessica Lares":
Noooooooo! ***Tom*** was the only reason why I even tuned into **TWiT**!

"MacNut"
More from **Leo**. He wants a full time in house news director. And has hired **Mike Elgan**.

"Rogifan"
Mike Elgan? Seriously?

"maflynn"
It sounds like that the ***Tom*** not working in the same studio was hampering the production of the show. **Skype** can only go so far, at least it seems for this situation.

I can't say this is a good or bad move until things shake out a bit more.
...

In figure 3.1 we see an example of ambiguity in the mention **Tom**. The ranker incorrectly links this mention to the Wikipedia page, *Tom_Friendly*, which is much more popular than the correct link, *Tom_Merritt*. Coherence fixes this mistake using the relational support of less ambiguous links like *TWiT.tv*.

An example from news text can be found in figure 3.2. Almost comically, the ranker links each mention *"Suleiman"* to the Wikipedia page of Suleiman the Magnificent, the 10th Sultan of the Ottoman Empire. Again, coherence is able to leverage information from other

Figure 3.2: Example of coherence in EL on a newswire data.

BEIRUT, Oct. 18 (Xinhua) – **Lebanon's** President **Michel Suleiman** said Friday that the international community was not sufficiently sharing the burden of displaced Syrians with **Lebanon**.
According to a statement by his media office, **Suleiman** told the ambassadors of major powers and the **United Nations Special Coordinator for Lebanon Derek Plumbly** that the Syrian refugees have put a severe strain on **Lebanon's** economy and "the participation of countries in sharing this burden is not sufficient."
...
Suleiman reiterated the need to establish UN-protected refugee camps inside Syria to facilitate the return of the displaced.
...

entities in the text to correctly link each of these ambiguous mentions to Michl Suleiman, the former president of Lebanon.

Each of these is a clear illustration of how coherence attempts to jointly ground all mentions in a text to set of entities which 'make sense' as a collection.

CHAPTER 4: CONCLUSIONS AND FUTURE WORK

In this paper we present what is, to the best of our knowledge, the first controlled study of the popular means of enforcing coherent assignments of entities in EL. We argue that many of the state of the art models for EL reduce to a similar architecture which means that coherence be applied modularly to improve any model. Furthermore, we argue that the coherence scoring function can be reduced to two primary components. Namely, the schema for representing candidates and the similarity measure which is used to quantify the similarity between two representations.

Breaking down the problem in this way makes it easy to see that candidate representation in coherence is an understudied problem, most models simply represent candidates as set of Wikipedia titles, e.g. outlinks or inlinks.

Much work ([3], [5], [18], [9]) show little improvement from incorporating coherence into their EL models. But we argue that this is a deficiency in execution rather than concept. Specifically, there has a been a keen lack of imagination in candidate representation.

Based on our experiments we conclude relations between entities are a powerful signal for understanding which candidate entities 'make sense' in a set. However, this information has been underutilized. Even the most simple binary model of 'do they relate or not?' shows improvement over the baseline in our experiments. In cases where relations have been used such as [5] and [18] the implementations either suffer from faulty assumptions about the quality of the ranker output, insufficient quantifications of relations or both.

A model that should be investigated further is that which was proposed by Wang et. al. [4] which is entirely a coherence model and which proposes more nuanced ways of quantifying relations between entities.

Conversely, more complex models for inference offered little gain in our experiments.

All of this points to the research question: how can we better represent entities for coherence? We believe this is more important than the similarity function which is what most research has emphasized. We believe that representation of candidates using structured data regarding relations is the way forward for research in entity linking with coherence.

REFERENCES

[1] R. Bunescu and M. Pasca, "Using encyclopedic knowledge for named entity disambiguation," in *Proceedings of the European Chapter of the ACL (EACL)*, 2006. [Online]. Available: http://www.cs.utexas.edu/ ml/publication/paper.cgi?paper=encyc-eacl-06.ps.gz

[2] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data." in *EMNLP 2007: Empirical Methods in Natural Language Processing, June 28-30, 2007, Prague, Czech Republic*, 01 2007, pp. 708–716.

[3] L. Ratinov and D. Roth, "Learning-based multi-sieve co-reference resolution with knowledge," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2012. [Online]. Available: http://cogcomp.cs.illinois.edu/papers/RatinovRo12.pdf

[4] S. Wiseman, A. M. Rush, S. M. Shieber, and J. Weston, "Learning anaphoricity and antecedent ranking features for coreference resolution," in *ACL*, 2015.

[5] X. Cheng and D. Roth, "Relational inference for wikification," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013. [Online]. Available: http://cogcomp.cs.illinois.edu/papers/ChengRo13.pdf

[6] C.-T. Tsai and D. Roth, "Cross-lingual wikification using multilingual embeddings," in *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 6 2016. [Online]. Available: http://cogcomp.cs.illinois.edu/papers/TsaiRo16b.pdf

[7] A. Globerson, N. Lazic, S. Chakrabarti, A. Subramanya, M. Ringaard, and F. Pereira, "Collective entity resolution with multi-focal attention," in *ACL*, 2016.

[8] V. I. Spitkovsky and A. X. Chang, "A cross-lingual dictionary for english wikipedia concepts." in *LREC*, 2012, pp. 3168–3175.

[9] N. Lazic, A. Subramanya, M. Ringgaard, and F. Pereira, "Plato: A selective context model for entity resolution," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 503–515, 2015. [Online]. Available: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/637

[10] H. Wang, J. Zheng, X. Ma, P. Fox, and H. Ji, "Language and domain independent entity linking with quantified collective validation." in *EMNLP*, 2015, pp. 695–704.

[11] X. Ling, S. Singh, and D. S. Weld, "Design challenges for entity linking," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 315–328, 2015. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/528

[12] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," in *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, 2008, pp. 25–30.

[13] R. Cilibrasi and P. M. B. Vitányi, "The google similarity distance," *CoRR*, vol. abs/cs/0412098, 2004.

[14] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[15] B. Hachey, J. Nothman, and W. Radford, "Cheap and easy entity evaluation." in *ACL*, 2014.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[17] K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon, "Representing text for joint embedding of text and knowledge bases." ACL Association for Computational Linguistics, September 2015.

[18] X. Ling, S. Singh, and D. S. Weld, "Design challenges for entity linking," *TACL*, vol. 3, pp. 315–328, 2015.