Entity Disambiguation with Web Links

Andrew Chisholm
School of Information Technologies
University of Sydney
NSW 2006, Australia
andy.chisholm.89@gmail.com

Ben Hachey
School of Information Technologies
University of Sydney
NSW 2006, Australia
ben.hachey@gmail.com

Abstract

Entity disambiguation with Wikipedia relies on structured information from redirect pages, article text, inter-article links, and categories. We explore whether web links can replace a curated encyclopaedia, obtaining entity prior, name, context, and coherence models from a corpus of web pages with links to Wikipedia. Experiments compare web link models to Wikipedia models on well-known CoNLL and TAC data sets.

Results show that using 34 million web links approaches Wikipedia performance. Combining web link and Wikipedia models produces the best-known disambiguation accuracy of 88.7 on standard newswire test data.

1 Introduction

Entity linking (EL) resolves mentions in text to their corresponding node in a knowledge base (KB), or NIL if the entity is not in the KB. Wikipedia and related semantic resources – Freebase, DBpedia, Yago2– have emerged as general repositories of notable entities. The availability of Wikipedia, in particular, has driven work on EL, knowledge base population (KBP), and semantic search. This literature demonstrates that the rich structure of Wikipedia– redirect pages, article text, inter-article links, categories – delivers disambiguation accuracy above 85% on newswire (He et al., 2013; Alhelbawy and Gaizauskas, 2014). But what disambiguation accuracy can we expect in the absence of Wikipedia's curated structure?

Web links provide much of the same information as Wikipedia inter-article links: anchors are used to derive alternative names and conditional probabilities of entities given names; in-link counts are used to derive a simple entity popularity measure; the text surrounding a link is used to derive textual context models; and overlap of in-link sources is used to derive entity cooccurrence models. On the other hand, web links lack analogues of additional Wikipedia structure commonly used for disambiguation, e.g., categories, encyclopaedic descriptions. Moreover, Wikipedia's editors ensure a clean and correct knowledge source while web links are a potentially noisier annotation source.

We explore linking with web links versus Wikipedia. Contributions include: (1) a new benchmark linker that instantiates entity prior probabilities, entity given name probabilities, entity context models, and efficient entity coherence models from Wikipedia-derived data sets; (2) an alternative linker that derives the same model using only alternative names and web pages that link to Wikipedia; (3) detailed development experiments, including analysis and profiling of Web link data, and a comparison of link and Wikipedia-derived models.

Results suggest that web link accuracy is at least 93% of a Wikipedia linker and that web links are complementary to Wikipedia, with the best scores coming from a combination. We argue that these results motivate open publishing of enterprise authorities and suggest that accumulating incoming links should be prioritised at least as highly as adding richer internal structure to an authority.

Transactions of the Association for Computational Linguistics, vol. 3, pp. 145–156, 2015. Action Editor: Ryan McDonald.
Submission batch: 10/2014; Revision batch 1/2015; Published 3/2015.

2 Related work

Thomas et al. (2014) describe a disambiguation approach that exploits news documents that have been curated by professional editors. In addition to consistently edited text, these include document-level tags for entities mentioned in the story. Tags are exploited to build textual mention context, assign weights to alternative names, and train a disambiguator. This leads to an estimated F_1 score of 78.0 for end-to-end linking to a KB of 32,000 companies. Our work is similar, but we replace quality curated news text with web pages and explore a larger KB of more than four million entites. In place of document-level entity tags, hyperlinks pointing to Wikipedia articles are used to build context, name and coherence models. This is a cheap form of third-party entity annotation with the potential for generalisation to any type of web-connected KB. However, it presents an additional challenge in coping with noise, including prose that lacks editorial oversight and links with anchor text that do not correspond to actual aliases.

Li et al. (2013) explore a similar task setting for microblogs, where short mention contexts exacerbate sparsity problems for underdeveloped entities. They address the problem by building a topic model based on Wikipedia mention link contexts. A bootstrapping approach analogous to query expansion augments the model using web pages returned from the Google search API. Results suggest that the bootstrapping process is beneficial, improving performance from approximately 81% to 87% accuracy. We demonstrate that adding link data leads to similar improvements.

The cold start task of the Text Analysis Conference is also comparable.[1] It evaluates how well systems perform end-to-end NIL detection, clustering and slot filling. Input includes a large document collection and a slot filling schema. Systems return a KB derived from the document collection that conforms to the schema. The evaluation target is long-tail or local knowledge. The motivation is the same as our setting, but we focus on cold-start linking rather than end-to-end KB population.

Finally, recent work addresses linking without and beyond Wikipedia. Jin et al. (2014) describe an unsupervised system for linking to a person KB from a social networking site, and Shen et al. (2014) describe a general approach for arbitrary KBs. Nakashole et al. (2013) and Hoffart et al. (2014) add a temporal dimension to NIL detection by focusing on discovering and typing emerging entities.

[1] http://www.nist.gov/tac/2014/KBP/ColdStart/guidelines.html

3 Tasks and art

Two evaluations in particular have driven comparative work on EL: the TAC KBP shared tasks and the Yago2 annotation of CoNLL 2003 NER data. We describe these tasks and their respective evaluation setup. A brief survey of results outlines the kind of performance we hope to achieve with link data. For task history, we suggest Hachey et al. (2013) and Shen et al. (2014). For an evaluation survey, see Hachey et al. (2014).

Our evaluation setup follows He et al. (2013) for comparability to their state-of-the-art disambiguation results across CoNLL and TAC data. Table 1 summarises the data sets used. Columns correspond to number of documents ($|\mathcal{D}|$), number of entities ($|\mathcal{E}|$), number of mentions ($|\mathcal{M}|$), and number of non-NIL mentions ($|\mathcal{M}_{\text{KB}}|$). The non-NIL mention number represents the set used for evaluation in the disambiguation experiments here. The table also includes average and standard deviation of the candidate set cardinality over $\mathcal{M}_{\text{KB}}$ ($\langle C \rangle$) and the percentage of mentions in $\mathcal{M}_{\text{KB}}$ where the correct resolution is in the candidate set ($R_{\mathcal{C}}$). The last column (SOA) gives the state-of-the-art score from the literature. Numbers are discussed below.

3.1 CoNLL

CoNLL is a corpus of Reuters newswire annotated for whole-document named entity recognition and disambiguation (Hoffart et al., 2011). CoNLL is public, free and much larger than most entity annotation data sets, making it an excellent evaluation target. It is based on the widely used NER data from the CoNLL 2003 shared task (Tjong Kim Sang and Meulder, 2003), building disambiguation on ground truth mentions. Training and development splits comprise 1,162 stories from 22-31 August 1996 and the held-out test split comprises 231 stories from 6-7 December 1996.

Data set	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{M}\rvert$	$\lvert\mathcal{M}_{\text{KB}}\rvert$ (%)	$\langle C\rangle$ (σ)	$R_{\mathcal{C}}$	SOA
CoNLL train	945	4,080	23,396	18,505 (79)	69 (194)	100	NA
CoNLL dev	216	1,644	5,917	4,791 (80)	73 (194)	100	79.7
CoNLL test	231	1,537	5,616	4,485 (80)	73 (171)	100	87.6
TAC train	1,040	456	1,500	1,070 (71)	23 (28)	94.4	NA
TAC test	1,012	387	2,250	1,017 (45)	24 (30)	88.5	81.0

Table 1: Data sets for disambiguation tasks addressed here. Statistics are described in Section 3.

The standard evaluation measure is *precision@1* ($p@1$) – the percentage of linkable mentions for which the system ranks the correct entity first (Hoffart et al., 2011). Linkable is defined as ground truth mentions for which the correct entity is a member of the candidate set. This factors out errors due to mention detection, coreference handling, and candidate generation, isolating the performance of the proposed ranking models. For comparability, we use Hoffart et al.'s Yago2 *means* relations for candidate generation. These alternative names are harvested from Wikipedia disambiguation pages, redirects and inter-article links. In the Hoffart et al. setting, candidate recall is 100%.

There are several key benchmark results for the CoNLL data set. Hoffart et al. (2011) define the task settings and report the first results. They employ a global graph-based coherence algorithm, leading to a score of 82.5. He et al. (2013) present the most comparable approach. Using deep neural networks, they learn entity representations based on similarity between link contexts and article text in Wikipedia. They report performance of 84.8 without collective inference, and 85.6 when integrating Han et al.'s (2011) coherence algorithm. Finally, Alhelbawy and Gaizauskas (2014) report the current best performance of 87.6 using a collective approach over a document-specific subgraph.

3.2 TAC 2010

Since 2009, the Text Analysis Conference (TAC) has hosted an annual EL shared task as part of its Knowledge Base Population track (KBP) (Ji and Grishman, 2011). Through 2013, the task is query-driven. Input includes a document and a name that appears in that document. Systems must output a KB identifier for each query, or NIL. The KB is derived from a subset of 818,741 Wikipedia articles. We use data from the 2010 shared task for several reasons. First, it facilitates comparison to current art. Second, it is a linking-only evaluation as opposed to linking plus NIL clustering. Finally, it includes comparable training and test data rather than relying on data from earlier years for training.

The TAC 2010 source collection includes news from various agencies and web log data. Training data includes a specially prepared set of 1,500 web queries. Test data includes 2,250 queries – 1,500 news and 750 web log uniformly distributed across person, organisation, and geo-political entities. Candidate generation here uses the DBpedia lexicalizations data set (Mendes et al., 2012), article titles, and redirect titles. We also add titles and redirects stripped of appositions indicated by a comma (e.g., `Montgomery, Alabama`) or opening round bracket (e.g., `Joe Morris (trumpeter)`). Candidate recall is 94.4 and 88.5 on the training and test sets – an upper limit on disambiguation accuracy.

Following He et al., we report KB accuracy (A_{KB}) - the percentage of correctly linked non-NIL mentions - to isolate disambiguation performance. Before evaluation, we map Wikipedia titles in our output to TAC KB identifiers using the Dalton and Dietz (2013) alignment updated with Wikipedia redirects. To our knowledge, Cucerzan (2011) report the best A_{KB} of 87.3 for an end-to-end TAC entity linking system, while He et al. (2013) report the best A_{KB} of 81.0 for a disambiguation-focused evaluation. There are a number of differences, e.g.: mention detection for coherence, coreference modelling, and substring matching in candidate generation. Analysis shows that these can have a large effect on system performance (Hachey et al., 2013; Piccinno and Ferragina, 2014). We use He et al.'s setup to control for differences and for comparability to He et al.'s results.

Component	Articles	Mentions	Web links
f_{prior}	68.4	68.4	63.0
f_{name}	69.2	69.2	58.4
f_{bow}	50.6	55.8	62.2
f_{dbow}	49.9	51.2	54.0

Table 2: p@1 results for individual components on the CoNLL development data. The first two columns correspond to the Wikipedia models described in Section 4.3, one derived from article text and the other from mention contexts. The last column corresponds to the web link models described in Section 5.

4 Wikipedia benchmark models

A wide range of EL approaches have been proposed that take advantage of the clean, well-edited information in Wikipedia. These include entity prior models derived from popularity metrics; alias models derived from Wikipedia redirects, disambiguation pages and inter-article links; textual context models derived from Wikipedia article text; and entity coherence models derived from the Wikipedia inter-article link graph. We survey these models and describe a new benchmark linker that instantiates them from existing Wikipedia-derived data sets. For a more detailed survey of features in supervised systems, see Meij et al. (2012) and Radford (2014).

Table 2 contains an overview of p@1 results for individual components on the CoNLL development data.

4.1 Entity prior

The simplest approach to entity disambiguation ranks candidate entities in terms of their popularity. For example, 0.000001% of inter-article links in Wikipedia point to `Nikola Tesla`, while 0.000008% point to `Tesla Motors`. An entity prior is used in generative models (Guo et al., 2009; Han and Sun, 2011) and in supervised systems that incorporate diverse features (Radford et al., 2012). We define the entity prior as the probability of a link pointing to entity e:

$$f_{prior}(e) = \log \frac{|\mathcal{I}_{*,e}|}{|\mathcal{I}_{*,*}|}$$

where $\mathcal{I}_{*,e} \in \mathcal{I}_{*,*}$ is the set of pages that link to entity e. We derive this from DBpedia's Wikipedia Pagelinks data set, which contains the link graph between Wikipedia pages.[2] Missing values are replaced with a small default log probability of -20, which works better than add-one smoothing in development experiments. On the CoNLL development data, entity prior alone achieves 68.4 p@1.

4.2 Name probability

Name probability models the relationship between a name and an entity. For example, 0.04% of links with the anchor text 'Tesla' point to `Nikola Tesla`, while 0.03% point to `Tesla Motors`. Name probability was introduced as an initial score in coherence-driven disambiguation (Milne and Witten, 2008), and is used in most state-of-the-art systems (Ferragina and Scaiella, 2010; Hoffart et al., 2011; Cucerzan, 2011; Radford et al., 2012). We define name probability as the conditional probability of a name referring to an entity:

$$f_{name}(e, n) = log \frac{|\mathcal{M}_{n,e}|}{|\mathcal{M}_{n,*}|}$$

where $\mathcal{M}_{n,e}$ is the set of mentions with name n that refer to entity e and $\mathcal{M}_{n,*}$ is all mentions with name n. We use existing conditional probability estimates from the DBpedia Lexicalizations data set (Mendes et al., 2012).[2] This derives mentions from Wikipedia inter-article links, where names come from anchor text and referent entities from link targets. Estimates for entities that have fewer than five incoming links are discarded. We smooth these estimates using add-one smoothing. On the CoNLL development data, name probability alone achieves 69.2 p@1.

4.3 Textual context

Textual context goes beyond intrinsic entity and name popularity, providing a means to distinguish between entities based on the words with which they occur. For example, references to `Tesla` the car manufacturer appear in passages with words like 'company', 'electric', 'vehicle'. References to the inventor appear with words like 'engineer', 'ac', 'electrical'. Textual context was the primary component of the top system in the first TAC evaluation (Varma et al., 2009), and is a key component in recent art (Ratinov et al., 2011; Radford et al., 2012).

[2] `http://wiki.dbpedia.org/Downloads`

BOW context We model textual context as a weighted bag of words (BOW), specifically as a term vector $\vec{t}$ containing $tfidf$ weights:

$$tfidf(t,p) = \sqrt{f(t,p)} \cdot \log\left(\frac{|\mathcal{D}|}{|\{d \in \mathcal{D} | t \in d\}|}\right)$$

where t is a term, p is a passage of text, $f(t,p)$ is the term frequency of t in p, $|\mathcal{D}|$ is the total number of documents, and $\{d \in \mathcal{D} | t \in d\}$ is the number of documents containing t (Salton and Buckley, 1988). We derive the term frequency for an entity e from the corresponding article content in the Kopiwiki plain text extraction (Pataki et al., 2012). Terms include three million token 1-3 grams from Mikolov et al. (2013), with the top 40 by document frequency as stop words. Candidate entities are scored using cosine distance between a mention context $\vec{t}_m$ and the entity model $\vec{t}_e$:

$$f_{bow}(m,e) = 1 - cos(\vec{t}_m, \vec{t}_e) = 1 - \frac{\vec{t}_m \cdot \vec{t}_e}{\|\vec{t}_m\| \|\vec{t}_e\|}$$

On the CONLL development data, BOW context derived from Wikipedia article text achieves 50.6 p@1. We also build entity models from their mention contexts, i.e., the combined text surrounding all incoming links. We project mentions into Kopiwiki article text, which yields more contexts than actual Wikipedia links. For an article a, we tag as mentions all aliases of entities linked to from a. We use aliases from Yago2 *means* relations (see Section 3.1). To ensure high precision, we only use aliases that are unambiguous with respect to the outlink set, have a length of at least two characters, include at least one upper-case character, and are not a member of the NLTK stop list. This is a noisy process, but gives us a pivot to assess whether differences observed later between Wikipedia and Web link models are due the way the context is modelled or the source of the context. The term frequency for an entity e is calculated over the concatenation of all contexts for e. BOW context derived from mentions achieves 55.8 p@1 on the CONLL development data, five points higher than article text.

DBOW context While BOW context models have been very successful, they require exact matching between terms and a large vocabulary. Distributional approaches model terms or concepts as semantic vectors (Pereira et al., 1993). Dimensionality reduction and deep learning improve generalisation and reduce vector size (Baroni et al., 2014). He et al. (He et al., 2013) report excellent performance using entity representations that optimise the similarity between mention contexts and article text in Wikipedia. However, this approach necessitates an expensive training process and significant run-time complexity. We introduce a simple distributed bag-of-words (DBOW) model that represents context as the $tfidf$-weighted average over word vectors $\mathcal{V}$:

$$\vec{v}_p = \frac{1}{|\mathcal{T}_p|} \sum_{t \in \mathcal{T}_p} tfidf(t,p) \cdot \vec{v}_t$$

where $\mathcal{T}_p$ is the set of terms in passage p, and $\vec{v}_t \in \mathcal{V}$ is the learnt word vector for term t. We use existing 300-dimensional word embeddings (Mikolov et al., 2013) and score candidates using cosine distance between mention context $\vec{v}_m$ and the entity model $\vec{v}_e$:

$$f_{dbow}(m,e) = 1 - cos(\vec{v}_m, \vec{v}_e)$$

On the CONLL development data, DBOW context models derived from article text and mention context achieve 49.9 and 51.2 respectively.

5 Web link models

The models above all have direct anologues in web links to Wikipedia articles. However, web links are a comparatively noisy source. For instance, anchors are less likely to be well-formed entity mentions, e.g., in links to `Semantic Web` we observe 'semantic markup' and 'Semantic Web Activity' as anchors. A lack of curation and quality control also allows for the misdirection of links. For example, we observe links to `Apple` the fruit where the surrounding context indicates an intention to link `Apple Inc` instead. It is an open question whether link-derived models are effective in disambiguation.

Below, we describe how models are instantiated using link data. We leverage the Wikilinks corpus of 9 million web pages containing a total of 34 million links to 1.7 million Wikipedia pages (Singh et al., 2012). This includes links to English Wikipedia pages that pass the following tests: (1) the page must not have >70% of sentences in common with a Wikipedia article; (2) the link must not be inside

	Wikipedia	Web links
Pages	8.7m	9.0m
Entities	8.9m	1.7m
Pairs	100.3m	31.2m

Table 3: Comparison of page-entity link graphs from Wikipedia and Wikilinks (in millions). These graphs are the basis for entity prior features (Sections 4.1, 5.1).

	Wikipedia	Web links
Names	1.4m	3.1m
Entities	1.5m	1.7m

Table 4: Comparison of name-entity link graphs from Wikipedia and Wikilinks (in millions). These graphs are the basis for name probability features (Sections 4.2, 5.2).

a table, near an image, or in obvious boilerplate material; (3) at least one token in the anchor text must match a token in the Wikipedia title; and (4) the anchor text must match a known alias from Wikipedia. The corpus provides the web page URL, the link anchor, and local textual content around each link.

Refer back to Table 2 for $p@1$ results for individual Web link components on the development data.

5.1 Entity prior

To instantiate f_{prior}, we build a page-entity link graph from Wikilinks. Where pages and entities are the same in the Wikipedia graph, here we have an unweighted bipartite graph of links from web pages to Wikipedia articles. On the CoNLL development data, the link-derived entity prior achieves 63.0 $p@1$. Table 3 characterises the two graphs. Note that the high entity count for Wikipedia here includes red links to articles that do not exist. The actual number of entities used in the Wikipedia model is 4.4 million. Nevertheless, while the two graphs have a similar number of pages that contain links, Wikipedia includes three times as many link pairs to 2.5 times as many entities. Furthermore, entities average 11.5 incoming links in the Wikipedia graph, compared to 3.5 in the Wikilinks graph. Nevertheless, the individual performance of the Web link prior is only 5.4 points shy of the corresponding Wikipedia prior.

Relative frequencies in Wikipedia and Wikilinks are similar, especially for entities that show up in the evaluation data. We observe a moderate correlation between entity priors from Wikipedia and Wikilinks ($\rho = 0.51$, $p < 0.01$), and a strong correlation across the subset of entities that occur in the development data ($\rho = 0.74$, $p < 0.01$).

5.2 Name probability

To instantiate f_{name}, we build a name-entity graph from Wikilinks. The structure is the same as the corresponding model from Wikipedia, both are bipartite graphs with cooccurrence frequencies on edges. However, names here are sourced from link anchors in web pages rather than Wikipedia articles. For comparability with the Wikipedia model, we ignore links to entities that occur fewer than five times. We observed no improvement using all links in development experiments. On the CoNLL development data, link-derived name probability achieves 58.4 $p@1$, more than ten points shy of the Wikipedia-derived name probability. Table 4 helps to explain this difference. Wikilinks has twice as many names linking to the same number of entities, resulting in more ambiguity and sparser models.

5.3 Textual context

To instantiate f_{bow} and f_{dbow}, we follow the same methodology used for Wikipedia mention contexts. The term frequency for an entity e is calculated over the concatenation of mention contexts for e. Document frequency is also calculated across aggregated entity contexts. Mention contexts include all text included in the Wikilinks data, a window of 46 tokens on average centred on the link anchor. Section 4.3 showed that Wikipedia mention contexts give better individual performance than Wikipedia article texts. Web link mentions result in even better performance. On the CoNLL development data, BOW context achieves 62.2 $p@1$, ten points higher than commonly used Wikipedia article model and seven points higher than the analogous Wikipedia mention model. DBOW context achieves 54.0 $p@1$, 2.8 points higher than the Wikipedia mention model.

Table 5 compares Wikipedia and Wikilinks coverage of entities from the CoNLL development set. The second column ($|\mathcal{E}|$) contains the number of unique entities that have usable context. Note that the entity universe we consider here is all article pages in English Wikipedia (4,418,901 total from the December 2013 Kopiwiki data set). The third

	$\lvert\mathcal{E}\rvert$	$Cov_{\mathcal{E}}$	$Cov_{\mathcal{M}}$	Joint
Articles	4,418,901	100	100	51.1
Mentions	954,698	77	89	58.3
Web links	1,704,703	82	92	64.1

Table 5: Coverage of textual context models for each source over entities ($\mathcal{E}$) and mentions ($\mathcal{M}$).

	$\bar{t}_{\mathcal{E}}$	$\bar{t}_{\mathcal{M}}$
Articles	438	438
Mentions	1653	50
Web links	922	46

Table 6: Mean in-vocab tokens per entity ($\bar{t}_{\mathcal{E}}$) and tokens per mention ($\bar{t}_{\mathcal{M}}$) for each textual context model.

and fourth columns correspond to coverage of entities ($Cov_{\mathcal{E}}$) and mentions ($Cov_{\mathcal{M}}$) from the CoNLL data set. Mention coverage exceeds entity coverage, highlighting the relationship with prevalence in newswire. The last column contains p@1 for the subset of mentions in CoNLL for which the correct resolution is covered by both articles and web links. This isolates context source, demonstrating that link contexts outperform article text.

Table 6 compares context size in Wikilinks to Wikipedia. Wikilinks BOW models are approximately twice the size of Wikipedia article models and half the size of Wikipedia mention models. This helps to explain why individual mention and link models outperform individual article models.

6 Learning to rank

To perform disambiguation, we first extract a set of real-valued features for each candidate entity e given a training set of mentions M. Features values are standardised to have zero mean and unit variance. Parameters of the training distribution are saved for consistent standardisation of test data.

We train a Support Vector Machine (SVM) classifier to perform pairwise ranking (Joachims, 2002). For each mention in the training set, we derive training instances by comparing the feature vector of the gold link ($\vec{f}_g$) with each non-gold candidate ($\vec{f}_c$):

$$(x_i, y_i) = \begin{cases} (\vec{f}_g - \vec{f}_c, +) & \text{if } i \text{ is odd} \\ (\vec{f}_c - \vec{f}_g, -) & \text{otherwise} \end{cases}$$

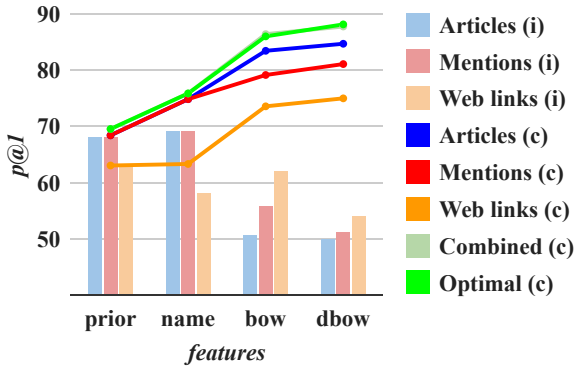

Figure 1: Individual (i) and cumulative (c) results for basic features on the CoNLL development data. Combined includes all features while Optimal includes the best subset. Optimal tracks Combined closely, but is just higher.

We create instances for the top-ten non-gold candidates by sum of absolute feature values:

$$activation(c) = \sum_{i=1}^{|\vec{f}_c|} |\vec{f}_{c,i}|.$$

In development experiments, this outperformed random selection and difference in activation. Class assignment is alternated to balance the training set.

To capture non-linear feature relationships we incorporate a degree-2 polynomial kernel via explicit feature mapping (Chang et al., 2010). Regularisation parameters are selected via grid search over the development set. Our final model utilises an L1 loss function, L2 weight penalty and C $\approx$ 0.03.

6.1 Feature selection

Sections 4 and 5 describe a total of ten model components, six from Wikipedia and four from Wikilinks. We select the optimal combination through exhaustive search. Figure 1 includes individual and cumulative results on the CoNLL development data. The article, mention and web link models each attain their best performance with all component features (entity, name, BOW, and DBOW): 84.7, 81.1, and 75.0 respectively. Adding mention context features doesn't improve the more conventional Wikipedia article model. Combining all features gives 87.7, while the optimal configuration achieves 88.1 without Wikipedia mention contexts. In the remaining experiments, optimal refers to Wikipedia article

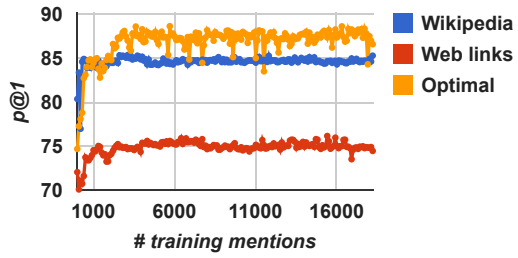

Figure 2: SVM learning curves for best configurations.

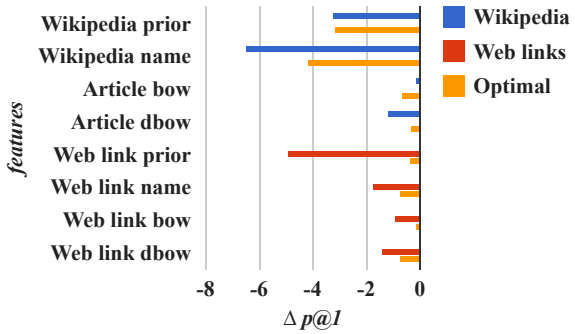

Figure 3: Ablation analysis of best configurations.

plus web link features and Wikipedia refers to article features alone.

6.2 Effect of training data size

Figure 2 compares learning curves for each model on CONLL development data. The x-axis corresponds to $p@1$ scores and the y-axis corresponds to the number of (randomly selected) mentions used in training. All models stabilise early, suggesting 6,000 annotated mentions are sufficient for the SVM to learn feature weights. Possibly due to higher quality and consistency of features, the Wikipedia model stabilises earlier, before 1,000 annotated mentions.

6.3 Ablation analysis

Figure 3 contains an ablation analysis for Wikipedia and Web link features, as well as the optimal overall combination of both. The most striking effect is due to the popularity components. Removing entity prior features reduces $p@1$ by 3.2 for Wikipedia and 5.0 for Web link. Removing name probability reduces $p@1$ by 6.5 for Wikipedia and 1.8 for Web link. In the overall model, the Wikipedia popularity components have a much larger impact (prior: -3.2, name: -4.2) than the Web link popularity components (prior: -0.4, name: -0.8). These results show the impact of noisy web links, which appears to be worse for name probability modelling. For context, removing DBOW features have a larger impact than BOW for both Wikipedia (BOW: -0.2, DBOW: -1.3) and Web link (BOW: -0.9, DBOW: -1.4). All individual context features have a small impact on the overall model despite redundancy.

7 Adding coherence

The model combinations above provide a strong, scalable baseline based on popularity and entity context. Another approach to context leverages the Wikipedia link graph to explicitly model the coherence among possible resolutions. Here, systems define some measure of entity-entity relatedness and maximise the coherence of entity assignments across the query document as a whole. This can be done using global methods over the entity link graph (Hoffart et al., 2011), but these have high runtime complexity. We employ a simple approach based on conditional probabilities:

$$p_{coh}(a|b) = \frac{|\mathcal{I}_a \cap \mathcal{I}_b|}{|\mathcal{I}_b|}$$

where $\mathcal{I}_e$ is the set of documents that link to entity e. The candidate-level feature is the average:

$$f_{cond}(e) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \log p_{coh}(e|c)$$

where $\mathcal{C}$ is the set of context entities for candidate entity e. For Wikipedia and Web link coherence, $\mathcal{I}_e$ models are derived respectively from the set of other articles that link to e and from the set of web pages that link to e. Given the same initial ranking from the optimal base model, Wikipedia and Web link coherence models alone achieve 84.7 and 76.6.

7.1 A two-stage classifier

To incorporate coherence, we use a two-stage classifier. First, we obtain an initial candidate ranking for each mention using the basic model described

	(a) CoNLL		(b) TAC 10	
	Pop	Ctx	Pop	Ctx
Wikipedia	**73.9**	53.3	72.6	65.0
Web links	62.5	**60.8**	73.3	**75.3**

Table 7: Web link components vs. Wikipedia.

	(a) CoNLL		(b) TAC 10	
	Base	+Coh	Base	+Coh
Wikipedia	**82.7**	**84.9**	78.6	80.2
Web links	77.0	80.7	78.5	80.2

Table 8: Web link combinations vs. Wikipedia.

	(a) CoNLL		(b) TAC 10	
	Base	+Coh	Base	+Coh
Wikipedia	82.7	84.9	78.6	80.2
+ Web links	**86.1**	**88.7**	79.6	80.7

Table 9: Web links complement Wikipedia.

in Section 6 above, and populate $\mathcal{C}$ from the top-one candidate for each unique context name. A second classifier incorporates all features, including basic components and coherence. Given the same initial ranking, adding coherence improves individual Wikipedia and Web link models 4.5 and 6.4 points to 89.2 and 81.4 $p@1$ on the CoNLL development data. These results suggests that coherence is a powerful feature to overcome low scores in the basic Web link model. But, coherence only improves the optimal combination of basic Wikipedia and web link features by 1.1 point to 89.2. This suggests coherence may not contribute much on top of an already strong set of basic features.

8 Final experiments

We report final experiments on the held-out CoNLL and TAC 2010 test sets. As described in Section 3 above, we report $p@1$ for CoNLL following Hoffart et al. (2011) and A_{KB} for TAC following He et al. (2013). We use a reference implementation to compute evaluation measures and pairwise significance (Hachey et al., 2014). We bold the superior configuration for each column only if the difference is significant ($p < 0.05$).

8.1 Results

Can link components replace KB components? Table 7 compares performance of basic model components. The popularity (Pop) column contains results using just entity prior and name probability features. The context (Ctx) column contains results using just BOW and DBOW features. Results follow trends observed in development experiments. Specifically, Wikipedia popularity models are better, but web link context models are better. Interestingy, web link popularity is significantly indistinguishable from Wikipedia popularity on TAC 10 data. This may be attributed to the fact that TAC selectively samples difficult mentions.

Can links replace a curated KB? Table 8 compares performance of the Wikipedia and Web link systems using the basic feature set alone and with coherence. Wikipedia models generally perform better. However, the Web link configurations perform at 93.1, 95.1, 99.9, and 100% of the Wikipedia linker – 97% on average. This suggests that a link data set can replace a curated KB, with only a small impact on accuracy. Results also show that adding coherence improves performance in all cases.

Do links complement article text? Table 9 compares a standard Wikipedia-only model to a model that also includes features derived from Web link data. Adding Web link data has a strong impact on CoNLL, improving both configurations by approximately 4 points. We observe less impact on TAC. Nevertheless, the large improvements on CoNLL provide good evidence for complementarity and recommend using both feature sets when available.

The state of the art Finally, Table 10 compares our Wikipedia and Web link combinations to state-of-the-art numbers from the literature. First, we note that adding coherence to our base model results in a significant improvement on CoNLL test data, but not on TAC 2010. For comparison the literature, we report 95% confidence intervals. If a confidence bar overlaps a reported number, the difference can not be assumed significant at $p < 0.05$. Results on TAC 10 are competitive with He et al.'s (2013) 81.0. On the CoNLL data, our best system achieves 88.7 $p@1$– a new state of the art. Furthermore, the best base model is competitive with previous art that uses complex collective approaches to coherence.

	DEV	CoNLL	TAC 10
Base model	87.7	86.1	79.6
- 95% CI	[85.3, 90.0]	[83.1, 88.8]	[77.1, 82.1]
Base+Coh	89.4	**88.7**	80.7
- 95% CI	[87.3, 91.2]	[86.2, 90.9]	[78.2, 83.1]
Hoffart	79.3	82.5	—
Houlsby	79.7	84.9	—
He	—	85.6	81.0
Alhelbawy	—	87.6	—

Table 10: Comparison to the disambiguation literature.

9 Discussion

We set out to determine whether links from external resources can replace a clean, curated KB. Wikipedia is an incredible resource that has advanced our understanding of and capabilities for identifying and resolving entity mentions. However, it covers only a small fraction of all entities. Applications that require other entities must therefore extend Wikipedia or use alternative KBs. We explore a setting where a custom KB is required, but it is possible to harvest external documents with links into the custom KB. Overall, results are promising for using links in a knowledge-poor setting. The link-derived system performs nearly as well as the rich-KB system on both of our held-out data sets.

Web link combinations perform at 97% of Wikipedia combinations on average. However, creating a KB as rich as Wikipedia represents an estimated 100 million hours of human effort (Shirky, 2010). We do not have a comparable estimate for the Web link data. However, it is created as byproduct of publishing activities and the labour pool is external. Considering this and the additional noise in web data, it is remarkable that the Web link models do so well with respect to the Wikipedia models.

We also present detailed experiments comparing popularity, context, and coherence components across settings. Here, results are even more surprising. As expected, Web link popularity and coherence models trail Wikipedia models. However, Web link context models outperform Wikipedia context models by 7 to 10 points.

We add the Web link components into the Wikipedia system to achieve, to our knowledge, the best published result of 88.7 on the CoNLL data set. Furthermore, results suggest that coherence modelling does not require complex global graph algorithms. Our simple approach improves performance over the basic model by one to three points. On the other hand, our basic system without coherence modelling approaches state-of-the-art performance on its own. This suggests that additional popularity and context features from web links can replace coherence where efficiency is a concern.

We believe these results have a number of implications for management of entity KBs. First, they motivate concerted efforts to link content to KBs since links lead to substantial accuracy improvements over a conventional model based on rich KB data alone. Second, it informs allocation of editorial resources between interlinking data sets and curating KBs. Since models built from link data alone approach state-of-the-art performance, curating links is a reasonable alternative to curating a KB. This is especially true if link curation is cheaper or if links can be created as a byproduct of other content authorship and management activities.

Finally, where KB data is currently proprietary, results here motivate openly publishing KB entities and encouraging their use as a disambiguation endpoint for public content. In addition to providing pathways to paid content, incoming links provide a simple means to harvest rich metadata from external content and this can be used to build high-quality resolution systems.

A key avenue for future work is to evaluate how well our approach generalises to other web KBs. For instance, incorporating links to sites like Freebase or IMDb which complement or extend Wikipedia's entity coverage.

10 Conclusion

Despite widespread use in entity linking, Wikipedia is clearly not the only source of entity information available on the web. We demonstrate the potential for web links to both complement and completely replace Wikipedia derived data in entity linking. This suggests that, given sufficient incoming links, any knowledge base may be used for entity linking. We argue that this motivates open publishing of enterprise KBs. Code is available under an MIT license at `https://github.com/wikilinks/nel`.

Acknowledgments

Andrew Chisholm is supported by a Google Faculty Research Award. Ben Hachey is the recipient of an Australian Research Council Discovery Early Career Researcher Award (DE120102900).

References

Ayman Alhelbawy and Robert Gaizauskas. 2014. Graph ranking for collective named entity disambiguation. In *Annual Meeting of the Association for Computational Linguistics*, pages 75–80.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Annual Meeting of the Association for Computational Linguistics*, pages 238–247.

Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. 2010. Training and testing low-degree polynomial data mappings via linear SVM. *Journal of Machine Learning Research*, 11:1471–1490.

Silviu Cucerzan. 2011. TAC entity linking by performing full-document entity extraction and disambiguation. In *Text Analysis Conference*.

Jeffrey Dalton and Laura Dietz. 2013. UMass CIIR at TAC KBP 2013 entity linking: query expansion using Urban Dictionary. In *Text Analysis Conference*.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *International Conference on Information and Knowledge Management*, pages 1625–1628.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *International Conference on Research and Development in Information Retrieval*, pages 267–274.

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

Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and easy entity evaluation. In *Annual Meeting of the Association for Computational Linguistics*, pages 464–469.

Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Annual Meeting of the Association for Computational Linguistics*, pages 945–954.

Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: a graph-based method. In *International Conference on Research and Development in Information Retrieval*, pages 765–774.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Annual Meeting of the Association for Computational Linguistics*, pages 30–34.

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 *Conference on Empirical Methods in Natural Language Processing*, pages 782–792.

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *International World Wide Web Conference*, pages 385–396.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Annual Meeting of the Association for Computational Linguistics*, pages 1148–1158.

Yuzhe Jin, Emre Kcman, Kuansan Wang, and Ricky Loynd. 2014. Entity linking at the tail: Sparse signals, unknown entities, and phrase models. In *International Conference on Web Search and Data Mining*, pages 453–462.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *International Conference on Knowledge Discovery and Data Mining*, pages 133–142.

Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. 2013. Mining evidences for named entity disambiguation. In *International Conference on Knowledge Discovery and Data Mining*, pages 1070–1078.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding semantics to microblog posts. In *International Conference on Web Search and Data Mining*, pages 563–572.

Pablo N. Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia: A multilingual cross-domain knowledge base. In *International Conference on Language Resources and Evaluation*, pages 1813–1817.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

David Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *Conference on Information and Knowledge Management*, pages 509–518.

Ndapandula Nakashole, Tomasz Tylenda, and Gerhard Weikum. 2013. Fine-grained semantic typing of

emerging entities. In *Annual Meeting of the Association for Computational Linguistics*, pages 1488–1497.

Máté Pataki, Miklós Vajna, and Attila Marosi. 2012. Wikipedia as text. *ERCIM News*, (89):48–49.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Annual Meeting of the Association for Computational Linguistics*, pages 183–190.

Francesco Piccinno and Paolo Ferragina. 2014. From TagME to WAT: a new entity annotator. In *SIGIR Workshop on Entity Recognition and Disambiguation*, pages 55–62.

Will Radford, Will Cannings, Andrew Naoum, Joel Nothman, Glen Pink, Daniel Tse, and James R. Curran. 2012. (Almost) Total Recall – SYDNEY_CMCRC at TAC 2012. In *Text Analysis Conference*.

Will Radford. 2014. *Named entity linking using rich knowledge*. Ph.D. thesis, The University of Sydney.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Annual Meeting of the Association for Computational Linguistics*, pages 1375–1384.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Wei Shan, Jiawei Han, and Jianyong Wang. 2014. A probabilistic model for linking named entities in web text with heterogeneous information networks. In *International Conference on Mangement of Data*, pages 1199–1210.

Wei Shen, Jianyon Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions (to appear). *Transactions on Knowledge and Data Engineering*.

Clay Shirky. 2010. *Cognitive surplus: Creativity and generosity in a connected age*. Allen Lane, London.

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.

Merine Thomas, Hiroko Bretz, Thomas Vacek, Ben Hachey, Sudhanshu Singh, and Frank Schilder. 2014. Newton: Building an authority-driven company tagging and resolution system (in press). In Emma Tonkin and Stephanie Taylor, editors, *Working With Text: Tools, techniques and approaches for text mining*. Chandos, Oxford, UK.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Conference On Computational Natural Language Learning*, pages 142–147.

Vasudeva Varma, Praveen Bysani, Kranthi Reddy, Vijay Bharat, Santosh GSK, Karuna Kumar, Sudheer Kovelamudi, Kiran Kumar N, and Nitin Maganti. 2009. IIIT Hyderabad at TAC 2009. In *Text Analysis Conference*.