Cheap and easy entity evaluation

Ben Hachey **Joel Nothman** **Will Radford**
School of Information Technologies
University of Sydney
NSW 2006, Australia
ben.hachey@sydney.edu.au
{joel,wradford}@it.usyd.edu.au

Abstract

The AIDA-YAGO dataset is a popular target for whole-document entity recognition and disambiguation, despite lacking a shared evaluation tool. We review evaluation regimens in the literature while comparing the output of three approaches, and identify research opportunities. This utilises our open, accessible evaluation tool. We exemplify a new paradigm of distributed, shared evaluation, in which evaluation software and standardised, versioned system outputs are provided online.

1 Introduction

Modern entity annotation systems detect mentions in text and disambiguate them to a knowledge base (KB). Disambiguation typically returns the corresponding Wikipedia page or NIL if none exists.

Named entity linking (NEL) work is driven by the TAC shared tasks on query-driven knowledge base population (Ji and Grishman, 2011). Evaluation focuses on disambiguating queried names and clustering NIL mentions, but most systems internally perform whole-document named entity recognition, coreference, and disambiguation (Cucerzan and Sil, 2013; Pink et al., 2013; Cheng et al., 2013; Fahrni et al., 2013). *Wikification* work generally evaluates end-to-end entity annotation including KB-driven mention spotting and disambiguation (Milne and Witten, 2008b; Kulkarni et al., 2009; Ratinov et al., 2011; Ferragina and Scaiella, 2010). Despite important differences in mention handling, NEL and wikification work have followed a similar trajectory. Yet to our knowledge, there are no comparative whole-document evaluations of NEL and wikification systems.

Public data sets have also driven research in whole-document entity disambiguation (Cucerzan, 2007; Milne and Witten, 2008b; Kulkarni et al., 2009; Bentivogli et al., 2010; Hoffart et al., 2011; Meij et al., 2012). However, with many task variants and evaluation methodologies proposed, it is very difficult to synthesise a clear picture of the state of the art.

We present an evaluation suite for named entity linking, leveraging and advocating for the AIDA disambiguation annotations (Hoffart et al., 2011) over the large and widely used CoNLL NER data (Tjong Kim Sang and Meulder, 2003). This builds on recent rationalisation and benchmarking work (Cornolti et al., 2013), adding an isolated evaluation of disambiguation. Contributions include:

- a simple, open-source evaluation suite for end-to-end, whole-document NEL;
- disambiguation evaluation facilitated by gold-standard mentions;
- reference outputs from state-of-the-art NEL and wikification systems published with the suite for easy comparison;
- implementation of statistical significance and error sub-type analysis, which are often lacking in entity linking evaluation;
- a venue for publishing benchmark results continuously, complementing the annual cycle of shared tasks;
- a repository for versioned corrections to ground truth annotation.

We see this repository, at https://github.com/wikilinks/conll03_nel_eval, as a model for the future of informal shared evaluation.

We survey entity annotation tasks and evaluation, proposing a core suite of metrics for end-to-end linking and tagging, and settings that isolate mention detection and disambiguation. A comparison of state-of-the-art NEL and wikification systems illustrates how key differences in mention handling affect performance. Analysis suggests that focusing evaluation too tightly on subtasks like candidate ranking can lead to results that do not reflect end-to-end performance.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 464–469,
Baltimore, Maryland, USA, June 23-25 2014.

2 Tasks and metrics

The literature includes many variants of the entity annotation task and even more evaluation approaches. Systems can be invoked under two settings: given text with expressions to be linked (gold mentions); or given plain text only (system mentions). The former enables a diagnostic evaluation of disambiguation, while the latter simulates a realistic end-to-end application setting.

Within each setting, metrics may consider different subsets of the gold ($\mathcal{G}$) and system ($\mathcal{S}$) annotations. Given sets of (doc, token span, kbid) tuples, we define precision, recall and F_1 score with respect to some annotation filter f:

$$P_f = \frac{|f(\mathcal{G}) \cap f(\mathcal{S})|}{|f(\mathcal{S})|} \quad , \quad R_f = \frac{|f(\mathcal{G}) \cap f(\mathcal{S})|}{|f(\mathcal{G})|}$$

We advocate two core metrics, corresponding to the major whole-document entity annotation tasks. *Link annotation* measures performance over every linked mention. Its filter f_L matches spans and link targets, disregarding NILs. This is particularly apt when entity annotation is a step in an information extraction pipeline. *Tag annotation* measures performance over document-level entity sets: f_T disregards span information and NILs. This is appropriate when entity annotation is used, e.g., for document indexing or social media mining (Mihalcea and Csomai, 2007; Meij et al., 2012). We proceed to ground these metrics and diagnostic variants in the literature.

2.1 End-to-end evaluation

We follow Cornolti et al. (2013) in evaluating end-to-end entity annotation, including both mention detection and disambiguation. In this context, f_L equates to Cornolti et al.'s *strong annotation match*; f_T measures what they call *entity match*.

2.2 Mention evaluation

Mention detection performance may be evaluated regardless of linking decisions. A filter f_M discards the link target (kbid). Of the present metrics, only this considers NIL-linked system mentions as different from non-mentions. For comparability with wikification, we consider an additional filter $f_{M_{\mathrm{KB}}}$ to NEL output that retains only linked mentions. f_M and $f_{M_{\mathrm{KB}}}$ are equivalent to Cucerzan's (2007) *mention evaluation* and Cornolti et al.'s *strong mention match* respectively. f_M is comparable to the NER evaluation from the CoNLL 2003 shared task (Tjong Kim Sang and Meulder, 2003): span equivalence is handled the same way, but metrics here ignore mention types.

2.3 Disambiguation evaluation

Most NEL and wikification literature focuses on disambiguation, evaluating the quality of link target annotations in isolation from NER error. Providing systems with ground truth mentions makes f_L equivalent to Mihalcea and Csomai's (2007) *sense disambiguation evaluation* and Milne and Witten's (2008b) *disambiguation evaluation*. It differs from Kulkarni et al.'s (2009) metric in being micro-averaged (equal weight to each mention), rather than macro-averaged across documents. f_L recall is comparable to TAC's KB *recall* (Ji and Grishman, 2011). It differs in that all mentions are evaluated rather than specific queries.

Related evaluations have also isolated disambiguation performance by: considering the links of only correctly identified mentions (Cucerzan, 2007); or only true mentions where the correct entity appears among top candidates before disambiguation (Ratinov et al., 2011; Hoffart et al., 2011; Pilz and Paass, 2012). We do not prefer this approach as it makes system comparison difficult. For comparability, we implement a filter $f_{L_{\mathrm{HOF}}}$ that retains only Hoffart-linkable mentions having a YAGO $means$ relation to the correct entity.

Tag annotation (f_T) with ground truth mentions is equivalent to Milne and Witten's (2008b) *link evaluation*, Mihalcea and Csomai's (2007) *keyword extraction evaluation* and Ratinov et al.'s (2011) *bag-of-titles evaluation*. It is comparable to Pilz and Paass's (2012) *bag-of-titles evaluation*, but does not account for sequential order and keeps all gold-standard links regardless of whether they are found by candidate generation.

2.4 Further diagnostics and rank evaluation

Several evaluations in the literature are beyond the scope of this paper but planned for future versions of the code. This includes further diagnostic subtask evaluation, particularly *candidate set recall* (Hachey et al., 2013), NIL *accuracy* (Ji and Grishman, 2011) and *weak mention matching* (Cornolti et al., 2013). With a score for each prediction, further metrics are possible: rank evaluation of tag annotation with *r-precision*, *mean reciprocal rank* and *mean average precision* (Meij et al., 2012); and rank evaluation of mentions for comparison to Hoffart et al. (2011) and Pilz and Paass (2012).

3 Data

The CoNLL-YAGO dataset (Hoffart et al., 2011) is an excellent target for end-to-end, whole-document entity annotation. It is public, free and much larger than most entity annotation data sets. 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. It has standard training and development splits that are representative of the held-out test data, all being sourced from the Reuters text categorisation corpus (Lewis et al., 2004), which is provided free for research purposes. Training and development comprise 1,162 stories from 22-31 August 1996 and held-out test comprises 231 stories from 6-7 December 1996. The layered annotation provides useful information for analysis including categorisation topics (e.g., general news, markets, sport) and NE type markup (PER, ORG, LOC, MISC).

The primary drawback is that KB annotations are currently present only if there is a YAGO *means* relation between the mention string and the correct entity. This means that there are a number of CoNLL entity mentions referring to entities that exist in Wikipedia that are nonetheless marked NIL in the ground truth (e.g. 'DSE' for 'Dhaka Stock Exchange'). This may be addressed by using a shared repository to adopt versioned improvements to the ground truth. Annotation over CoNLL tokenisation sometimes results in strange mentions (e.g., 'Washington-based' instead of 'Washington'). However, prescribed tokenisation simplifies comparison and analysis.

Another concern is that link annotation goes stale, since Wikipedia titles are only canonical with respect to a particular point in time. This is because pages may be renamed or reorganised:

- to improve editorial structure, such as downgrading an entity from having a page of its own, to a mere section in another page;
- to account for newly notable entities, such as creating a disambiguation page for a title that formerly had a single known referent; or
- because of changes in fact, such as corporate mergers and name changes.

All systems compared provide Wikipedia titles as labels, which are mapped to current titles for comparison: for each entity title t linked in the gold data, we query the Wikipedia API to find t's canonical form t_c and retrieve titles of all redirects to t_c.

4 Reference systems

Even on public data sets, comparison to published results can be very difficult and extremely costly (Fokkens et al., 2013). We include reference system output in our repository for simple comparison. Other researchers are welcome to add reference output, providing a continuous benchmark that complements the annual cycle of large shared tasks like TAC KBP.

4.1 TagMe

TagMe (Ferragina and Scaiella, 2010) is an end-to-end wikification system specialising in short texts. TagMe performs best among publicly available wikification systems (Cornolti et al., 2013). Mention detection uses a dictionary of anchor text from links between Wikipedia pages. Candidate ranking is based on entity relatedness (Milne and Witten, 2008a), followed by mention pruning. We use thresholds on annotation scores supplied by Marco Cornolti (personal communication) of 0.289 and 0.336 respectively for mention/link and tag evaluation. TagMe annotations may not align with CoNLL token boundaries, e.g., <annot title="Oakland, New Jersey">OAKLAND, N.J</annot>. Before evaluation, we extend annotations to overlapping tokens.

4.2 AIDA

AIDA (Hoffart et al., 2011) is the system presented with the CoNLL-YAGO dataset and places emphasis on state-of-the-art ranking of candidate entity sets. Mentions are ground truth from the CoNLL data to isolate ranking performance, equivalent to applying the $f_{L_{\text{HOF}}}$ filter. Ranking is informed by a graph model of entity compatibility.

4.3 Schwa

Schwa (Radford et al., 2012) is a heuristic NEL system based on a TAC 2012 shared task entrant. Mention detection uses a NER model trained on news text followed by rule-based coreference. Disambiguation uses an unweighted combination of KB statistics, document compatibility (Cucerzan, 2007), graph similarity and targeted textual similarity. Candidates that score below a threshold learned from TAC data are linked to NIL. The system is very competitive, performing at 93% and 97% respectively of the best accuracy numbers we know of on 2011 and 2012 TAC evaluation data (Cucerzan and Sil, 2013).

System	Mentions	Filter	P	R	F_1
Cucerzan	System	f_M	82.2	84.8	83.5
Schwa	System	f_M	86.9	76.7	81.5
TagMe	System	$f_{M_{\text{KB}}}$	75.2	60.4	67.0
Schwa	System	$f_{M_{\text{KB}}}$	82.5	74.5	78.3

Table 1: Mention detection results. Cucerzan results as reported (Cucerzan, 2007).

5 Results

We briefly report results over the reference systems to highlight characteristics of the evaluation metrics and task settings. Results hinge upon Schwa since we have obtained only its output in all settings. Except where noted, all differences are significant ($p < 0.05$) according to approximate randomisation (Noreen, 1989), permuting annotations over whole documents.

5.1 Mention evaluation

Table 1 evaluates mentions with and without NILs. None of the systems reported use a CoNLL-trained NER tagger, for which top shared task participants approached 90% F_1 in a stricter evaluation than f_M. We note the impressive numbers reported by Cucerzan (2007) using a novel approach to mention detection based on capitalisation and corpus co-occurrence statistics, and the similar performance[1] to Schwa, whose NER component is trained on another news corpus.

In wikification, NIL-linked mentions may not be relevant, and it may suffice to identify only the most canonical forms of names, rather than all mentions in a coreference chain. With $f_{M_{\text{KB}}}$, Schwa has much higher recall than TagMe, though TagMe's precision is understated because it generates non-NE annotations that are not present in the CoNLL-YAGO ground truth (e.g., linking 'striker' to `Forward (association football)`).

5.2 Disambiguation evaluation

Table 2 contains results isolating disambiguation performance. AIDA ranking outperforms Schwa according to both the link ($f_{L_{\text{HOF}}}$) and tag metrics ($f_{T_{\text{HOF}}}$). If we remove the Hoffart et al. (2011) linkable constraint, we observe that Schwa disambiguation performance loses about 8 points in precision on the link metric (f_L) and 2 points on the tag metric (f_T). This suggests that disambiguation evaluation without the linkable constraint is important, especially if the application requires detecting and disambiguating all mentions.

System	Mentions	Filter	P	R	F_1
Schwa	Gold	f_L	67.5	78.3	72.5
Schwa	Gold	$f_{L_{\text{HOF}}}$	79.7	78.3	79.0
AIDA	Gold	$f_{L_{\text{HOF}}}$	83.2	83.2	83.2
Schwa	Gold	f_T	77.8	77.7	77.7
Schwa	Gold	$f_{T_{\text{HOF}}}$	80.1	77.6	78.8
AIDA	Gold	$f_{T_{\text{HOF}}}$	87.7	84.2	85.9

Table 2: Disambiguation results for mention-level linking and document-level tagging.

System	Mentions	Filter	P	R	F_1
TagMe	System	f_L	63.2	50.7	56.3
Schwa	System	f_L	67.6	61.0	64.2
TagMe	System	f_T	65.0	65.4	65.2
Schwa	System	f_T	71.2	62.6	66.6

Table 3: End-to-end results for mention-level linking and document-level tagging.

The comparison here highlights a notable evaluation intricacy. The Schwa system disambiguates all gold mentions rather than those with KB links, and the document compatibility approach means that evidence from a NIL mention may offer confounding evidence when linking linkable mentions. Further, although using the same mentions, systems use search resources with different recall characteristics, so the Schwa system may not retrieve the correct candidate to disambiguate.

5.3 End-to-end evaluation

Finally, Table 3 contains end-to-end entity annotation results. Again, these results highlight key differences in mention handling between NEL and wikification. Coreference modelling helps NEL detect and link ambiguous names (e.g., 'President Bush') that refer to the same entity as unambiguous names in the same text (e.g., 'George W. Bush'). And restricting the the universe to named entities is appropriate for the CoNLL-YAGO data. The advantage is marked in the mention-level link evaluation (f_L). However, the systems are statistically indistinguishable in the document-level tag evaluation (f_T). Thus the extra NER and coreference machinery may not be justified if the application is document indexing or social media mining (Meij et al., 2012), wherein a KB-driven mention detector may be favourable for other reasons.

[1] Significance cannot be tested since we do not have the Cucerzan (2007) output.

Error	$f_{L_{\text{HOF}}}$		f_L	
	AIDA	Schwa	TagMe	Schwa
wrong link	752	896	429	605
link as nil	-	79	-	111
nil as link	-	-	183	337
missing	-	-	1,780	1,031
extra	-	-	1,663	927

Table 4: $f_{L_{\text{HOF}}}$ and f_L error profiles.

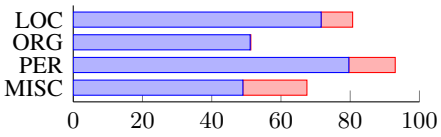

Figure 1: Schwa f_L and $f_{L_{\text{HOF}}}$ F_1 for NE types

6 Analysis

We analyse the types of error that a system makes. We also harness the multi-layered annotation to quantify the effect of NE type and document topic.

By error type Table 4 shows error counts based on the disambiguation link evaluation with the linkable constraint ($f_{L_{\text{HOF}}}$) and the end-to-end link evaluation (f_L). Errors are divided as follows:

wrong link: mention linked to wrong KB entry
link as nil: KB-entity mention linked to NIL
nil as link: NIL mention linked to the KB
missing: true mention not detected
extra: mention detected spuriously

AIDA outperforms Schwa under the linkable evaluation, making fewer wrong link errors. Schwa also overgenerates NIL, which may reflect candidate recall errors or a conservative disambiguation threshold. On the end-to-end evaluation, Schwa makes more linking errors (wrong link, link as nil, nil as link) than TagMe, but fewer in mention detection, leading to higher overall performance.

By entity type Figure 1 evaluates only mentions where the CoNLL 2003 corpus (Tjong Kim Sang and Meulder, 2003) marks a NE mention of each type. This is based on the link evaluation of Schwa. The left and right bars correspond to end-to-end (f_L) and disambiguation ($f_{L_{\text{HOF}}}$) F_1 respectively. In accord with TAC results (Ji and Grishman, 2011), high accuracy can be achieved on PER when a full name is given, while ORG is substantially more challenging. MISC entities are somewhat difficult to disambiguate, with identification errors hampering end-to-end performance.

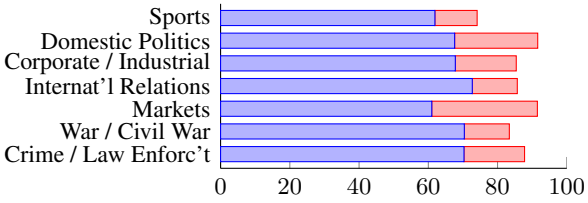

Figure 2: Schwa f_L and $f_{L_{\text{HOF}}}$ F_1 for top topics

By topical category The underlying Reuters Corpus documents are labelled with topic, country and industry codes (Lewis et al., 2004). Figure 2 reports F_1 on test documents from each frequent topic. It highlights that much ambiguity remains unresolved in *Sports*, while very high performance linking is attainable in categories such as *Markets* and *Domestic Politics*, only when given ground truth linkable mentions.

7 Conclusion

We surveyed entity annotation tasks and advocated a core set of metrics for mention, disambiguation and end-to-end evaluation. This enabled a direct comparison of state-of-the-art NEL and wikification systems, highlighting the effect of key differences. In particular, NER and coreference modules make NEL approaches suitable for applications that require all mentions, including ambiguous names and entities that are not in the KB. For applications where document-level entity tags are appropriate, the NEL and wikification approaches we evaluate have similar performance.

The big picture we wish to convey is a new approach to community evaluation that makes benchmarking and qualitative comparison cheap and easy. In addition to the code being open source, we use the repository to store reference system output, and – we hope – emendations to the ground truth. We encourage other researchers to contribute reference output and hope that this will provide a continuous benchmark to complement the current cycle of shared tasks.

Acknowledgements

Many thanks to Johannes Hoffart, Marco Cornolti, Xiao Ling and Edgar Meij for reference outputs and guidance. Ben Hachey is the recipient of an Australian Research Council Discovery Early Career Researcher Award (DE120102900). The other authors were supported by the Capital Markets CRC Computable News project.

References

Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta, and Kateryna Tymoshenko. 2010. Extending English ACE 2005 corpus annotation with ground-truth links to Wikipedia. In *COLING Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27.

Xiao Cheng, Bingling Chen, Rajhans Samdani, Kai-Wei Chang, Zhiye Fei, Mark Sammons, John Wieting, Subhro Roy, Chizheng Wang, and Dan Roth. 2013. Illinois cognitive computation group UI-CCG TAC 2013 entity linking and slot filler validation systems. In *Text Analysis Conference*.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *22nd International Conference on the World Wide Web*, pages 249–260.

Silviu Cucerzan and Avirup Sil. 2013. The MSR systems for entity linking and temporal slot filling at TAC 2013. In *Text Analysis Conference*.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716.

Angela Fahrni, Benjamin Heinzerling, Thierry Göckel, and Michael Strube. 2013. HITS' monolingual and cross-lingual entity linking system at TAC 2013. In *Text Analysis Conference*.

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

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 1691–1701.

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

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.

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

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *15th International Conference on Knowledge Discovery and Data Mining*, pages 457–466.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding semantics to microblog posts. In *5th International Conference on Web Search and Data Mining*, pages 563–572.

Rada Mihalcea and Andras Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In *16th Conference on Information and Knowledge Management*, pages 233–242.

David Milne and Ian H. Witten. 2008a. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *AAAI Workshop on Wikipedia and Articial Intelligence*, pages 25–30.

David Milne and Ian H. Witten. 2008b. Learning to link with Wikipedia. In *17th Conference on Information and Knowledge Management*, pages 509–518.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses*. John Wiley & Sons.

Anja Pilz and Gerhard Paass. 2012. Collective search for concept disambiguation. In *24th International Conference on Computational Linguistics*, pages 2243–2258.

Glen Pink, Will Radford, Will Cannings, Andrew Naoum, Joel Nothman, Daniel Tse, and James R. Curran. 2013. SYDNEY_CMCRC at TAC 2013. In *Text Analysis Conference*.

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*.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *49th Annual Meeting of the Association for Computational Linguistics*, pages 1375–1384.

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.