Combining Open Source Annotators for Entity Linking through Weighted Voting

Pablo Ruiz and **Thierry Poibeau**
LATTICE Lab
CNRS, École Normale Supérieure, U Paris 3 Sorbonne Nouvelle
1, rue Maurice Arnoux, 92120 Montrouge, France
{pablo.ruiz.fabo,thierry.poibeau}@ens.fr

Abstract

An English entity linking (EL) workflow is presented, which combines the annotations of five public open source EL services. The annotations are combined through a weighted voting scheme inspired by the ROVER method, which had not been previously tested on EL outputs. The combined results improved over each individual system's results, as evaluated on four different golden sets.

1 Introduction

The Entity Linking (EL) literature has shown that the quality of EL systems' results varies widely depending on characteristic of the corpora they are applied to, or on the types of entities we need to link (Cornolti et al., 2013, Usbeck et al., 2015). For instance, a system that links to a wide set of entity types can be less accurate at basic types like *Person*, *Location*, *Organization* than systems specializing in those basic types.

A way to make up for the uneven performance of entity linking methods across corpora would be mixing different annotators' results, so that the annotators' strengths complement each other. This paper presents a method to combine the outputs of five open source entity linking systems, in order to obtain improved results. The method involves a weighted voting scheme that had not been previously applied to EL, and improves annotation results across four test-corpora.

The structure of the paper is as follows: Section 2 presents related work. Section 3 describes the combined entity linking system. Section 4 provides an evaluation of the system's results and a discussion.

2 Related Work

General surveys on EL can be found in (Cornolti et al., 2013) and (Rao et al., 2013). Besides the EL literature, work on combining NLP annotators is particularly relevant for the present article.

The goal of combining different NLP systems is obtaining combined results that are better than the results of each individual system. Fiscus (1997) created the ROVER method, with weighted voting to improve speech recognition outputs. ROVER was found to improve parsing results by De la Clergerie et al. (2008). In Named Entity Recognition (NER), Rizzo et al. (2014) improved results combining systems via different machine learning algorithms.

In entity linking, the potential benefits of combining annotations have been explored before. Rizzo and Troncy (2012) describe the NERD system, which combines entity linkers. However, we are not aware of a system that, like ours, makes an automatic choice among the systems' conflicting annotations, based on an estimate of each annotation's quality. Our approach to choose among conflicting annotations is inspired by the ROVER method, which had not been previously attempted for EL to our knowledge. A further difference in our system is that the set of linkers we combine is public and open-source.

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 211–215,
Denver, Colorado, June 4–5, 2015.

3 Combining Annotators

Our workflow performs English EL to Wikipedia, combining the outputs of the following EL systems: Tagme 2[1] (Ferragina and Scaiella, 2010), DBpedia Spotlight[2] (Mendes et al. 2011), Wikipedia Miner[3] (Milne and Witten, 2008a), AIDA[4] (Hoffart et al., 2011) and Babelfy[5] (Moro et al. 2014). A description of the different systems can be found in (Usbeck et al., 2015). The systems rely on a variety on algorithms and it can be expected that their results will complement each other.

3.1 Obtaining Individual Annotator Outputs

First of all, a client requests the annotations for a text from each linker's web-service, using the services' default settings except for the confidence threshold,[6] which is configured in our workflow.

We obtained optimal thresholds for each system (Column *t* in Tables 1 and 2) with the BAT Framework[7] (Cornolti et al., 2013). The BAT Framework allows calling several entity linking tools and compares their results using different annotation confidence thresholds, with a view to finding the thresholds that yield best results according to several evaluation measures.

Annotations are filtered out if their confidence is below the thresholds obtained in the way just described. The remaining annotations proceed to the annotation-voting step.

3.2 Pre-ranking Annotators

Our annotation voting exploits annotators' precision on an annotated reference set in order to weight the annotations produced by each annotator (details in 3.3 below). It is not viable to create a reference set for each new corpus that we need to perform entity linking on. To help overcome this issue, we adopt the following approach: We have ranked the annotators for precision on two reference sets: AIDA/CONLL Test B (Hoffart et al., 2011), and IITB (Kulkarni et al., 2009). The IITB dataset contains annotations for category *Others*, i.e. entities that are not a person, organization or location, whereas AIDA/CONLL B does not contain such annotations. The proportion of annotations in a corpus that fall into the *Others* category is a strong predictor of annotators' performance on that corpus, according to a study on how different dataset features correlate with annotators' results, available on the GERBIL platform[8] (Usbeck et al., 2015). Taking this into account, in order to annotate a new corpus, if annotations for the *Others* category are needed for that new corpus, the annotator ranking for the IITB corpus will be used in order to weight the new corpus' annotations, since IITB is the only one among our two reference sets that contains annotations for *Others*, and an annotator performing well on IITB is likely to perform well when annotations for *Others* are needed. If, conversely, annotations for the *Others* category are not needed, the annotator ranking for the AIDA/CONLL B reference corpus is used in order to weight the new corpus' annotations.

3.3 Annotation Voting Scheme

The voting scheme is in Figure 1. Each annotation is formalized as a pairing between a mention *m* (a span of characters in the text) and a Wikipedia entity *e*. For each annotation <*m*, *e*>, Ω_m is the set of annotations whose mentions overlap[9] with *m*. The set Ω_m is divided into disjoint subsets, each of which contains annotations linking to a different entity. Each subset *L* is voted by *vote*(*L*): For each annotation *o* in *L*, *N* is the number of annotators we combine (i.e. 5), $r_{o,anr}$, is the rank of annotator *anr*, which produced annotation *o*, and P_{anr} is *anr*'s precision on the ranking reference corpus (see 3.2 above). Finally, parameter α influences the distance between the annotations' votes based on their annotators' rank, and was set at 1.75 based on the best results on both ranking reference corpora.

[1] http://tagme.di.unipi.it/tagme_help.html

[2] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

[3] http://wikipedia-miner.cms.waikato.ac.nz/

[4] https://github.com/yago-naga/aida

[5] http://babelfy.org/download.jsp

[6] The public deployments were used, but for AIDA, which was set up locally: Source v2.1.1, Data 2010-08-17v7. In AIDA, the tech=GRAPH option was used (non-default, but recommended by AIDA's authors for benchmarking).

[7] https://github.com/marcocor/bat-framework

[8] See *Annotator - Dataset feature correlations* at http://gerbil.aksw.org/gerbil/overview

[9] Assume two mentions (p1, e1) and (p2, e2), where p1 and p2 are the mentions' first character indices, and e1 and e2 are the mentions' last character indices. The mentions overlap iff $((p1 = p2) \wedge (e1 = e2)) \vee ((p1 = p2) \wedge (e1 < e2)) \vee ((p1 = p2) \wedge (e2 < e1)) \vee ((e1 = e2) \wedge (p1 < p2)) \vee ((e1 = e2) \wedge (p2 < p1)) \vee ((p1 < p2) \wedge (p2 < e1)) \vee ((p2 < p1) \wedge (p1 < e2))$.

```
for each set $\Omega_m$ of overlapping annotations:
  for $L \in \Omega_m$:
      $vote(L) = \dfrac{\sum_{o \in L} \left(N - \left(r_{o,anr} - \alpha\right)\right) \cdot P_{o,anr}}{N}$
  if $\max_{L \in \Omega_m} ( vote(L) ) > P_{max}$ : select $\mathrm{argmax}_{L \in \Omega_m}(vote(L))$
```

Figure 1: Entity voting scheme.

The entity for the subset L which obtains the highest vote among Ω_m's subsets is selected if its vote is higher than P_{max}, i.e. the maximum precision for all annotators on the ranking corpus.[10] Once an entity has been selected for a set of overlapping mentions, the mention itself needs to be selected. Best results were obtained when the most common mention in the set was selected. In case of ties, the longest mention among the most common ones was selected (e.g. if two mentions occur twice each in the set, select the longer one).

4 Evaluation and Results

4.1 Evaluation Method

Datasets: The workflow was tested on four golden sets. First, the two datasets that had also been used as reference sets in order to obtain the weights to vote annotations with (see Section 3.2). These two datasets were AIDA/CONLL B (231 documents with 4485 annotations; 1039 characters avg., news and sports topics) and IITB (103 documents with 11245 annotations; 3879 characters avg., topics from news, science and others). In order to test whether the annotator weights obtained from those two corpora can improve results when applied to annotator combination on other corpora, we tested on two additional datasets: MSNBC (Cucerzan, 2007), with 20 documents and 658 annotations (3316 characters avg., news topics) and AQUAINT (Milne and Witten, 2008b), with 50 documents and 727 annotations (1415 characters avg., news topics).

The AQUAINT dataset contains annotations for common noun entities (besides Person, Location, Organization). For this reason, according to the procedure described in 3.2 above, its annotations were weighted according to annotators' ranking on the IITB corpus, which also contains common-noun annotations. The MSNBC dataset does not contain common-noun annotations, so the annotator ranking for the AIDA/CONLL test-set was used in order to combine annotations in MSNBC.

Measures: The EL literature has stressed the importance of evaluating systems on more than one measure. We tested the workflow on strong annotation match (SAM) and entity match (ENT) (Cornolti et al., 2013). SAM requires an annotation's position to exactly match the reference, besides requiring the entity annotated to match the reference entity. ENT ignores positions and only evaluates whether the entity proposed by the system matches the reference.

Mapping files: Evaluating EL to Wikipedia requires making sure that we consider the same set of target entities for each EL system, since the versions of Wikipedia deployed within each system may differ. A mapping between current Wikipedia titles for the golden set annotations and non-canonical forms for these titles was created (including e.g. older titles redirecting to the new ones), and applied to golden and system sets before evaluation.[11]

Tools: Evaluation was carried out with the *neleval* tool[12] from the TAC-KBP Entity Discovery and Linking task (Ji et al., 2014). The tool implements several EL-relevant metrics, accepting a common delimited format for golden sets and results across corpora. The tool's significance testing function via randomized permutation/bootstrap methods was also applied to our results.

4.2 Results and Discussion

Results are provided in Table 1 (SAM measure) and Table 2 (ENT measure). Note that, to promote transparency, individual system annotations, combined results, reference annotations and mapping files are available on a website.[13] Each table shows micro-averaged precision, recall and F1 on the four golden sets, for each individual system, plus results for the combined workflow in the last row. The optimal confidence thresholds for each annotator are also indicated where applicable.

[10] See Table 1 and Table 2 below for P_{max} values in the ranking reference corpora: P_{max} is the maximum (excluding row *Combined*) in columns AIDA/CONLL B and IITB.

[11] The mapping was created based on *fetch_map* from the conll03_nel_eval tool by Hachey et al. (2013), https://github.com/wikilinks/conll03_nel_eval

[12] https://github.com/wikilinks/neleval/wiki

[13] https://sites.google.com/site/entitylinking1/

Corpus	AIDA/CONLL B				IITB				MSNBC				AQUAINT			
System	*t*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1
Tagme	0.219	54.8	53.9	54.4	0.086	41.1	42.6	41.8	0.188	44.7	42.4	*43.5*	0.188	39.9	46.5	43.0
Spotlight	0.086	28.1	38.8	32.6	0.016	41.0	48.2	44.3	0.063	21.8	28.1	24.6	0.055	15.6	45.3	23.2
W Miner	0.57	45.3	50.3	47.7	0.25	55.2	44.4	*49.2*	0.664	42.3	38.2	40.2	0.57	34.8	57.6	*43.4*
AIDA	0.0	76.7	46.7	*58.1*	0.0	50.2	5.6	10.0	0.0	63.6	23.8	34.7	0.0	50.3	27.7	35.7
Babelfy	dna	34.7	34.0	34.3	dna	46.8	14.9	22.7	dna	31.8	28.8	31.1	dna	22.6	31.5	26.3
Combined	dna	64.8	61.7	***61.9**	dna	59.3	44.7	***50.0**	dna	54.3	43.4	***48.2**	dna	34.1	64.1	**44.5**

Table 1: **Strong annotation match (SAM).** Optimal confidence thresholds (*t*), Micro-averaged Precision, Recall, F1 for each annotator and combined system. Babelfy and the combined system use no confidence thresholds (dna).

Corpus	AIDA/CONLL B				IITB				MSNBC				AQUAINT			
System	*t*	P	R	F1	*T*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1
Tagme	0.234	58.2	67.9	62.7	0.102	47.6	45.7	46.7	0.328	66.8	49.9	57.1	0.198	63.8	55.4	59.3
Spotlight	0.094	30.8	40.1	34.8	0.008	36.6	51.8	42.9	0.063	21.6	27.5	24.2	0.055	26.2	49.8	34.3
W Miner	0.477	46.9	57.3	51.6	0.195	61.3	43.3	*50.6*	0.664	50.1	52.8	51.4	0.523	59.9	62.5	*61.1*
AIDA	0.0	79.7	79.7	***79.7**	0.0	61.4	11.72	19.7	0.0	74.6	56.3	*64.2*	0.0	67.8	37.3	48.1
Babelfy	dna	35.6	37.9	36.7	dna	48.4	16.3	24.4	dna	36.5	37.5	37.0	dna	39.1	37.8	38.3
Combined	dna	65.0	78.5	*71.1*	dna	60.7	44.6	***51.4**	dna	66.7	62.3	**64.4**	dna	58.4	67.3	***62.5**

Table 2: **Entity match (ENT).** Optimal confidence thresholds (*t*), Micro-averaged Precision, Recall, F1 in for each annotator and combined system. Babelfy and the combined system use no confidence thresholds (dna).

The annotator rankings and weights with which annotations were weighted in our voting scheme (Figure 1) can be read off the *P* column for the ranking reference corpora (AIDA/CONLL or IITB). For instance, results for MSNBC were combined using the ranking from AIDA/CONLL. In terms of Figure 1, this means that MSBC annotations (for the SAM measure) were weighted with the following values, in format (Annotator, Rank, Weight): (AIDA, 0, 0.767), (Tagme, 1, 0.548), (Wikipedia Miner, 2, 0.453), (Babelfy, 3, 0.347), (Spotlight, 4, 0.281). The P_{max} value that each annotation's vote is compared to in MSNBC is 0.767.

In the tables, the best F1 score in each corpus is marked in bold, and the second-best F1 is in italics. The combined workflow obtains the best score in all cases, except ENT scores on AIDA/CONLL B. For the SAM measure, the improvements range between 0.8 points and 4.7 points of F1. For the ENT measure, improvements range between 0.2 and 1.4 points of F1. The differences are statistically significant in the majority of cases (scores with a star). Significance ($p < 0.05$) was assessed with the random permutation method in the *neleval* tool[12].

The combined workflow was able to improve over the best individual system regardless of which this system was: Tagme, Wikipedia Miner or AIDA. In some cases, the improvements over the best individual system's F1 take place because of markedly increased recall in the combined system compared to the best individual system's recall, without a major decrease in precision in the combined system (see AQUAINT results for ENT). The opposite pattern of improvement is also attested: In the MSNBC results for SAM, it is the increased precision of the combined workflow that makes its F1 improve over the best individual system's F1.

Regarding the significant drop in F1 in the combined system vs. the best individual system (AIDA) in the ENT results for the AIDA/CONLL B corpus, note that, in this case, the difference between AIDA's individual results and the results for the second-best individual system was much higher (17.2 points of F1) than anywhere else in the rest of tests performed. When such a large difference exists between the best individual system and the rest, an alternative type of voting may be needed in order to improve results over the best individual system.

5 Conclusion and Future Work

A workflow that combines the outputs of public open source entity linking (EL) systems via weighted voting was presented. The simple voting scheme generally improved F1 scores over the best individual system's F1, as assessed by the strong annotation match and entity match measures. Besides some enhancements to the voting scheme, interesting future work could be comparing this simple scheme's results with a more complex combination method, e.g. involving supervised learning based on available corpora annotated for entity linking (with mention–entity pairings).

Acknowledgements

Pablo Ruiz was supported through a PhD scholarship from Région Île de France.

References

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. (2013). A framework for benchmarking entity-annotation systems. In *Proc. of WWW*, 249–260.

Silviu Cucerzan. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proc. EMNLP and CNLL*, 708–716.

Éric V. De La Clergerie, Olivier Hamon, Djamel Mostefa, Christelle Ayache, Patrick Paroubek, and Anne Vilnat. (2008). Passage: from French parser evaluation to large sized treebank. In *Proc. of LREC 2008*, 3570–3576.

Paolo Ferragina and Ugo Scaiella. (2010). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proc. of CIKM'10*, 1625–1628.

Jonathan G. Fiscus. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, 1997*, 347–354.

Ben Hachey, Joel Nothman, and Will Radford. (2014). Cheap and easy entity evaluation. In *Proc. ACL*, 464–469.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. (2011). Robust disambiguation of named entities in text. In *Proc. of EMNLP*, 782–792.

Heng Ji, Joel Nothman and Ben Hachey. (2014). Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In *Proc. Text Analysis Conference*.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. (2009). Collective annotation of Wikipedia entities in web text. In *Proc. ACM SIGKDD*, 457–466.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. (2011). DBpedia spotlight: shedding light on the web of documents. In *Proc. of the 7th Int. Conf. on Semantic Systems, I-SEMANTICS'11*, 1–8.

David Milne and Ian H. Witten. (2008a). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proc. of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy,* 25–30.

David Milne and Ian H. Witten. (2008b). Learning to link with Wikipedia. In *Proc. CIKM*, 509–518.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. (2014). Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the ACL*, 2, 231–244.

Delip Rao, Paul McNamee, and Mark Dredze. (2013). Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*, 93–115. Springer.

Giuseppe Rizzo and Raphaël Troncy. (2012). NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proc. of the Demonstrations at EACL'12*, 73–76.

Giuseppe Rizzo, Marieke van Erp, and Raphaël Troncy. (2014). Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In *Proc. of LREC 2014*, 4593–4600.

Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga, Ciro Baron, Andrea Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccino, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. (2015). GERBIL–General Entity Annotator Benchmarking Framework. In *Proc. of WWW*.