ELCO3: Entity Linking with Corpus Coherence Combining Open Source Annotators

Pablo Ruiz, Thierry Poibeau, Frédérique Mélanie-Becquet

ELCO3: Entity Linking with Corpus Coherence Combining Open Source Annotators

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

Abstract

Entity Linking (EL) systems' performance is uneven across corpora or depending on entity types. To help overcome this issue, we propose an EL workflow that combines the outputs of several open source EL systems, and selects annotations via weighted voting. The results are displayed on a UI that allows the users to navigate the corpus and to evaluate annotation quality based on several metrics.

1 Introduction

The Entity Linking (EL) literature has shown that the performance of EL systems varies widely depending on the corpora they are applied to and of the types of entities considered (Cornolti et al., 2013). For instance, a system linking to a wide set of entity types can be less accurate at basic types like *Organization*, *Person*, *Location* than systems specializing in those basic types. These issues make it difficult for users to choose an optimal EL system for their corpora.

To help overcome these difficulties, we have created a workflow whereby entities can be linked to Wikipedia via a combination of the results of several existing open source EL systems. The outputs of the different systems are weighted according to how well they performed on corpora similar to the user's corpus.

Our target users are social science researchers, who need to apply EL in order to, for instance, create entity co-occurrence network visualizations. These researchers need to make informed choices about which entities to include in their analyses, and our tool provides metrics to facilitate these choices.

The paper is structured as follows: Section 2 describes related work. Section 3 presents the different steps in the workflow, and Section 4 focuses on the steps presented in the demo.

2 Related work

Cornolti et al. (2013) provide a general survey on EL. Work on combining EL systems and on helping users select a set of linked entities to navigate a corpus is specifically relevant to our workflow. Systems that combine entity linkers exist, e.g. NERD (Rizzo et al., 2012). However, there are two important differences in our workflow. First, the set of entity linkers we combine is entirely open source and public. Second, we use a simple voting scheme to optionally offer automatically chosen annotations when linkers provide conflicting outputs. This type of weighted vote had not previously been attempted for EL outputs to our knowledge, and is inspired on the ROVER method (Fiscus, 1997, De la Clergerie et al., 2008).

Regarding systems that help users navigate a corpus by choosing a representative set of linked entities, our reference is the ANTA tool (Venturini and Guido, 2012).[1] This tool helps users choose entities via an assessment of their corpus frequency and document frequency. Our tool provides such information, besides a measure of each entity's coherence with the rest of entities in the corpus.

[1] https://github.com/medialab/ANTA

3 Workflow description

The user's corpus is first annotated by making requests to three EL systems' web services: Tagme 2[2] (Ferragina and Scaiella, 2010), DBpedia Spotlight[3] (Mendes et al. 2011) and Wikipedia Miner[4] (Milne and Witten, 2008). Annotations are filtered out if their confidence score is below the optimal thresholds for those services, reported in Cornolti et al. (2013) and verified using the BAT-Framework.[5]

3.1 Annotation voting

The purpose of combining several linkers' results is obtaining combined annotations that are more accurate than each of the linkers' individual results. To select among the different linkers' outputs, a vote is performed on the annotations that remain after the initial filtering described above.

Our voting scheme is based on De la Clergerie et al.'s (2008) version of the ROVER method. An implementation was evaluated in (Ruiz and Poibeau, 2015). Two factors that our voting scheme considers are annotation confidence, and the number of linkers having produced an annotation. An important factor is also the performance of the annotator having produced each annotation on a corpus similar to the user's corpus: At the outset of the workflow, the user's corpus is compared to a set of reference corpora along dimensions that affect EL results, e.g. text-length or lexical cohesion[6] in the corpus' documents. Annotators that perform better on the reference corpus that is most similar along those dimensions to the user's corpus are given more weight in the vote.

In sum, the vote helps to select among conflicting annotation candidates, besides helping identify unreliable annotations.

3.2 Entity types

Entity types are assigned by exploiting information provided in the linkers' responses, e.g. DBpedia ontology types or Wikipedia category labels. The entity types currently assigned are *Organization*, *Person*, *Location*, *Concept*.

3.3 Entity coherence measures

Once entity selection is completed, a score that quantifies an entity's coherence with the rest of entities in the corpus is computed. This notion of coherence consists of two components. The first one is an entity's relatedness to other entities in terms of Milne and Witten's (2008) Wikipedia Link-based Measure (WLM, details below). The second component is the distance between entities' categories in a Wikipedia category graph.

WLM scores were obtained with Wikipedia Miner's *compare* method for Wikipedia entity IDs.[7] WLM evaluates the relatedness of two Wikipedia pages as a function of the number of Wikipedia pages linking to both, and the number of pages linking to each separately. In the literature, WLM has been exploited to disambiguate among competing entity senses within a document, taking into account each sense's relatedness to each of the possible senses for the remaining entity-mentions in the document. We adopt this idea to assess entity relatedness at corpus level rather than at document level. To do so, we obtain each entity's averaged WLM relatedness to the most representative entities in the corpus. The most representative entities in the corpus were defined as a top percentage of the entities, sorted by decreasing annotation confidence, whose annotation frequency and confidence are above given thresholds.

The second component of our entity coherence measure is based on distance between nodes in a Wikipedia category graph (see Strube and Ponzetto, 2006 for a review of similar methods). Based on the category graph, the averaged shortest path[8] between an entity and the most representative entities (see criteria above) of the same type was computed. Some categories like "People from {City}*"* were ignored, since they created spurious connections.

3.4 Annotation attributes

The final annotations contain information like position (document, character and sentence), confi-

[2] http://tagme.di.unipi.it/tagme_help.html
[3] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki
[4] http://wikipedia-miner.cms.waikato.ac.nz/
[5] https://github.com/marcocor/bat-framework
[6] Our notion of lexical cohesion relies on token overlap across consecutive token sequences, inspired on the block comparison method from Hearst (1997).
[7] http://wikipedia-miner.cms.waikato.ac.nz/services/?compare
[8] Using igraph.GraphBase.get_all_shortest_paths from the Python interface to igraph: http://igraph.org/python/

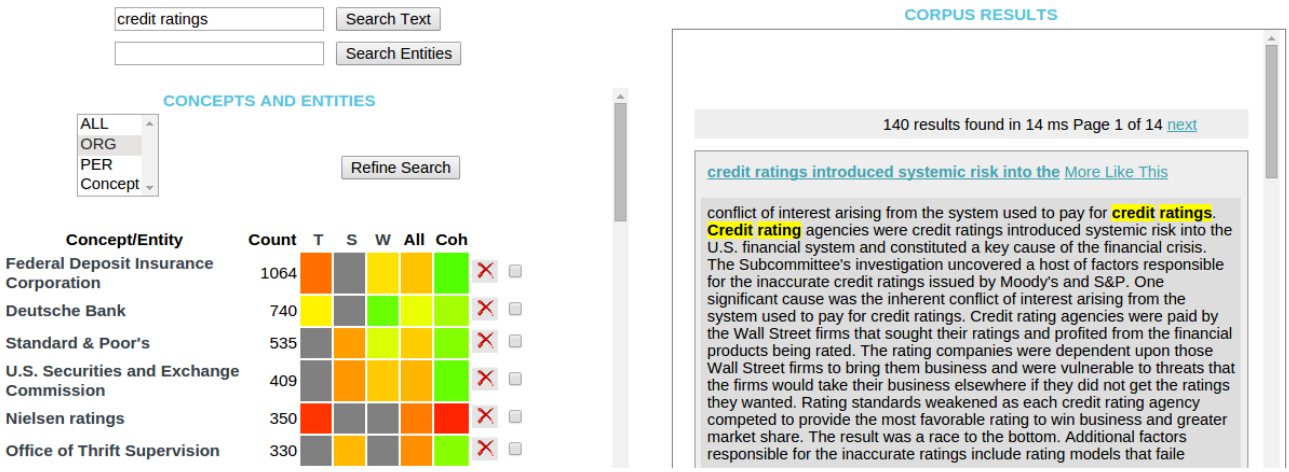

Figure 1: Results for query *credit ratings*. The right panel shows documents matching the query; the left panel shows the entities that have been annotated in those documents.

dence, and entity-type. This can be exploited for further textual analyses, e.g. co-occurrence networks.

4 Demonstrator

The goal of the workflow is to help users choose a representative set of entities to model a corpus, with the help of descriptive statistics and other measures like annotation confidence, or the coherence scores described above. A practical way to access this information is a UI, where users can assess the validity of an entity by simultaneously looking at its metrics, and at the documents where that entity was annotated. We present an early stage prototype of such a UI, which shows some of the features of the workflow above, using preprocessed content—the possibility to tag a new corpus is not online.

The demo interface[9] allows to navigate a corpus through search and entity facets. In Figure 1, a *Search Text* query displays, on the right panel, the documents matching the query,[10] while the entities annotated in those documents are shown in the left panel. A *Search Entities* query displays the entities matching the query on the left panel, and, on the right, the documents where those entities were annotated. *Refine Search* restricts the results on the right panel to documents containing certain entities or entity types, if the corresponding checkboxes at the end of each entity row, or items on the entity-types list have been selected. The colors provide a visual indication of the entity's confidence for each linker (columns, *T*, *S*, *W*, *All*), scaled[11] to a range between 0 (red) and 1 (green). Hovering over the table reveals the scores in each cell.

For the prototype, the corpus was indexed in Solr[12] and the annotations were stored in a MySQL DB. The EL workflow was implemented in Python and the UI is in PHP.

Examples of the utility of the information on the UI and of the workflow's outputs follow.

Usage example 1: Spotting incorrect annotations related to a search term. The demo corpus is about the 2008 financial crisis. Suppose the user is interested in organizations appearing in texts that mention *credit ratings* (Figure 1). Several relevant organizations are returned for documents matching the query, but also an incorrect one: *Nielsen ratings*. This entity is related to *ratings* in the sense of audience ratings, not credit ratings. The coherence score (column *Coh*) for the incorrect entity is much lower (red, dark) than the scores for the relevant entities (green, light). The score helps to visually identify the incorrect annotation, based on its lack of coherence with representative entities in the corpus.

Figure 1 also gives an indication how the different linkers complement each other: Some annotations have been missed by one linker (grey cells), but the other two provide the annotation.

[9] http://129.199.228.10/nav/gui/

[10] The application's Solr search server requires access to traffic on port 8983. A *connection refused* (or similar) error message in the search results panel is likely due to traffic blocked on that port at the user's network.

[11] scikit-learn: sklearn.preprocessing.MinMaxScaler.html

[12] http://lucene.apache.org/solr/

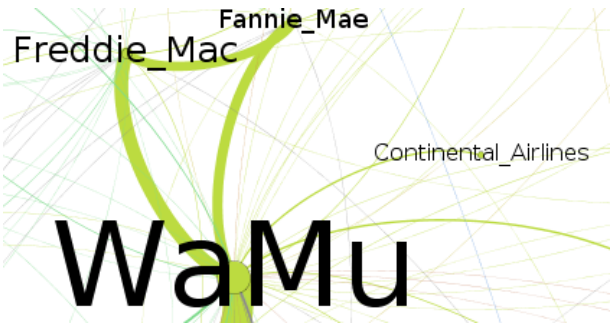

Figure 2: Region of an entity network created outside of our workflow, based on the individual output of one of the EL systems we combine. Node *Continental Airlines* in the network is an error made by that EL system.

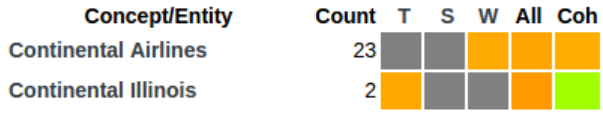

Figure 3: Result of a search in our GUI for entity labels containing *Continental*. The lower coherence score (*Coh*) for *Continental Airlines* (orange, dark) vs. *Continental Illinois* (green, light) suggests that the latter is correct and that the airline annotation is an error.

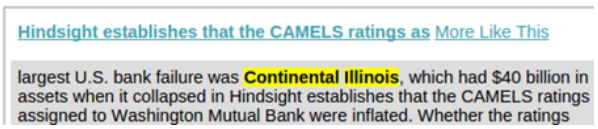

Figure 4: Example document showing that *Continental Illinois* is the correct entity in the corpus

Usage example 2: Verifying correctness of entities in networks. A common application of EL is creating co-occurrence networks, e.g. based on an automatic selection of entities above a certain frequency. This can result in errors. Figure 2 shows a small area from an entity co-occurrence network for our corpus. Our corpus comes from the 2014 PoliInformatics challenge (Smith et al., 2014), and the corpus topic is the 2008 financial crisis. The network was created independently of the workflow described in this paper, using Gephi,[13] based on entities annotated by Wikipedia Miner, which is one of the EL systems whose outputs our workflow combines. Node *Continental Airlines* in the network seems odd for the corpus, in the sense that the corpus is about the financial crisis, and Continental Airlines was not a major actor in the crisis. A *Search Entities* query for *Continental* on our GUI returns two annotations (Figure 3): the airline, and *Continental Illinois* (a defunct bank). The coherence (*Coh*) score for the bank is higher than for the airline. If we run a *Search Text* query for *Continental* on our GUI, the documents returned for the query confirm that the correct entity for the corpus is the bank (Figure 4 shows one of the documents returned).

The example just discussed also shows that the coherence scores can provide information that is not redundant with respect to annotation frequency or annotation confidence. It is the bank's coherence score that suggests its correctness: The incorrect annotation (for the airline) is more frequent, and the confidence scores for both annotations are equivalent.

In short, this second example is another indication how our workflow helps spot errors made by annotation systems and decide among conflicting annotations.

A final remark about entity networks: Our workflow segments documents into sentences, which would allow to create co-occurrence networks at sentence level. Some example networks based on our outputs and created with Gephi are available on the demo site.[14] These networks were not created programmatically from the workflow: The current implementation does not automatically call a visualization tool to create networks, but this is future work that would be useful for our target users.

5 Conclusion

Since entity linking (EL) systems' results vary widely according to the corpora and to the annotation types needed by the user, we present a workflow that combines different EL systems' results, so that the systems complement each other. Conflicting annotations are resolved by a voting scheme which had not previously been attempted for EL. Besides an automatic entity selection, a measure of coherence helps users decide on the validity of an annotation. The workflow's results are presented on a UI that allows navigating a corpus using text-search and entity facets. The UI helps users assess annotations via the measures displayed and via access to the corpus documents.

[13] http://gephi.github.io

[14] Follow link *Charts* on http://129.199.228.10/nav/gui

Acknowledgements

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

References

Cornolti, M., Ferragina, P., & Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. In *Proc. of WWW*, 249–260.

De La Clergerie, É. V., Hamon, O., Mostefa, D., Ayache, C., Paroubek, P., & Vilnat, A. (2008). Passage: from French parser evaluation to large sized treebank. In *Proc. LREC 2008*, 3570–3576.

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

Fiscus, J. G. (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*, 347–354.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics, 23*(1), 33–64.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. In *Proc. I-SEMANTICS'11*, 1–8.

Milne, D. & Witten, I. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proc. of AAAI Workshop on Wikipedia and Artificial Intelligence,* 25–30

Rizzo, G., & Troncy, R. (2012). NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proc. of the Demonstrations at EACL'13*, 73–76.

Ruiz, P. and Poibeau, T. (2015). Combining Open Source Annotators for Entity Linking through Weighted Voting. In *Proceedings of *SEM. Fourth Joint Conference on Lexical and Computational Semantics.* Denver, U.S.

Venturini, T. and Daniele Guido. 2012. Once upon a text: an ANT tale in Text Analytics. *Sociologica*, 3:1-17. Il Mulino, Bologna.

Smith, N. A., Cardie, C., Washington, A. L., Wilkerson, J. D. (2014). Overview of the 2014 NLP Unshared Task in PoliInformatics. *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*, 5–7.

Strube, M. and Ponzetto, S. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI*, vol. 6, 1419–1424.