# UBC Entity Discovery and Linking & Diagnostic Entity Linking at TAC-KBP 2014

**Ander Barrena, Eneko Agirre, Aitor Soroa**
IXA NLP Group / University of the Basque Country, Donostia, Basque Country
ander.barrena@ehu.es, e.agirre@ehu.es, a.soroa@ehu.es

## Abstract

This paper describe the runs submitted by the UBC team at TAC-KBP 2014 for both English Entity Discovery and Linking (EDL) and Diagnostic Entity Linking (DEL) tasks. Our main interest was to compare the performance between two totally different name entity recognizer systems and to combine them with three different name entity disambiguation systems that were developed for the TAC-KBP 2013 EL task. Therefore, we tried 6 possible detection-disambiguation combinations for EDL task. The results show that all system combinations attain similar scores, and that the best result is obtained by combining a supervised name entity recognizer with a random forest classifier for disambiguation. For the DEL task our best performance was obtained by disambiguating mentions with a Personalized PageRank algorithm. All systems reported to both tracks reached at least top 10, and also, scored between best and median performance in all the cases.

## 1 Introduction

This year the TAC-KBP 2014 organizers introduced a new task called Entity Discovery and Linking. The task consists in detecting all name mentions of a given document, and linking them to a reference entity of a specific Knowledge Base (KB). Mention of entities that are not in the reference Knowledge Base are linked to NIL. It is also necessary to classify the entities as person (*PER*), organization (*ORG*) or geopolitical entity (*GPE*). Finally, mentions linked to NIL have to be clustered, so that all mentions referring to the same NIL entity are in the same cluster.

Our approach for EDL consists of five steps. First, given a document we detect all possible entity mentions. Then, we generate the possible Wikipedia candidates for each mention. Thirdly, we disambiguate the mention by ranking the candidates entities and choosing the highest ranked one. We also assign a type (*PER*, *ORG* or *GPE*) to the highest ranked entity. Finally, we cluster the NIL mentions. Our system relies on the disambiguation step for both classifying and NIL clustering.

The DEL task mimics previous TAC-KBP contests and, therefore, mentions are given by the organizers. Besides, it is not necessary to assign a entity type. We use the same candidate generation, disambiguation and clustering algorithms for both EDL and DEL tasks.

We reused candidate generation and disambiguation algorithms tested in previous year at TAC-KBP English Entity Linking 2013 (Barrena, Agirre and Soroa, 2014), introducing minor changes in text preprocessing and the NIL clustering step.

## 2 Resources

We use a 2011 Wikipedia snapshot in our experiments. From the snapshot we extract two information resources: a dictionary and textual contexts for all candidate entities.

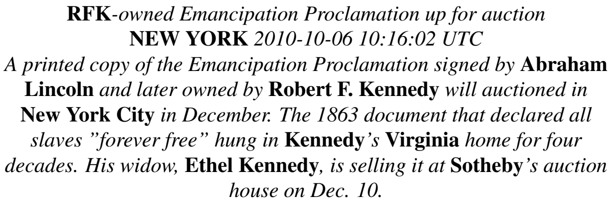
The dictionary is an association between strings and Wikipedia articles. We construct the dictionary using article titles, redirections, disambiguation pages, and anchor text. Mentions are lowercased and all text between parenthesis is removed. If the mention links to a disambiguation page, it is associated with all possible articles the disambiguation page points to. Each association between a string and article is scored with the prior probability, estimated as the number of times that the mention occurs in the anchor text of an article divided by the total number of occurrences of the mention. Note that our dictionary can disambiguate any mention, just returning the article with highest score.

We also extract textual contexts for all the possible candidate entities . Given an entity, we collect all the mentions to this entity within Wikipedia, and extract a context of 50 words around the anchor link. Contexts are lemmatized and POS tagged using the Stanford CoreNLPsuite[1].

In order to classify entities, we built a entity-type resource by gathering all classified named entities from the DBpedia and Yago2 ontologies. DBpedia uses information extracted from Wikipedia infoboxes and the ontology contains about 4 million instances. We selected all instances under the Place, Organization or Person ontological categories according to the DBpedia ontology[2]. Yago2[3] is a large Knowledge Base derived form Wikipedia, WordNet and GeoNames, which comprises more than 10 million entities. As Yago2 entities are mapped to WordNet, we select those entities which fall into person, organization and location according to WordNet. Merging this two information resources we create a new resource where each named entity is linked to the corresponding category.

## 3 Mention Detection

In order to detect mentions for the EDL task we used two different methods. The first method, called *Match-up*, recognizes mentions in the text when they fulfill the following two conditions: a) they occur as anchor texts in Wikipedia, and b) they contain some uppercased character. The second method, called *Ixa-pipe-nerc*, is a traditional Named Entity Recognition and Classification tagger trained on the CoNLL 2003 corpus (Agerri, Bermudez and Rigau, 2014). It is part of IXA pipes[4], a multilingual NLP pipeline developed by the IXA NLP Group. We only consider a mention if it is recognized as a Named Entity of type *PER*, *GPE* or *ORG*.



**RFK***-owned Emancipation Proclamation up for auction*
**NEW YORK** *2010-10-06 10:16:02 UTC*
*A printed copy of the Emancipation Proclamation signed by* **Abraham Lincoln** *and later owned by* **Robert F. Kennedy** *will auctioned in* **New York City** *in December. The 1863 document that declared all slaves "forever free" hung in* **Kennedy***'s* **Virginia** *home for four decades. His widow,* **Ethel Kennedy***, is selling it at* **Sotheby***'s auction house on Dec. 10.*

Figure 1: Example of a document from train data where bold marks gold mentions, those that mention detection systems should detect.

One of our main interests is to test how those two different name detection systems perform, and the impact of NER systems in final result. For example, figure 1 shows a sample from train data document [5], where gold mentions are marked in bold. For example, mentions like "NEW YORK", "Virginia", "Abraham Lincoln", "Sotheby" and "Robert F. Kennedy" are detected by both systems. However "RFK" it is only detected by *Match-up*, but produces false positives like "UTC", "Emancipation Proclamation", "December", "Dec" because they fulfill both conditions mentioned before. We deal with those mentions discarding them because they do not refer to a person, location or geo-political entity (see section 6).

So the question is, it is worthwhile to over generate in order to get the maximun recall, and discard later?

## 4 Candidate Ranking

We used three different Named Entity Disambiguation (NED) algorithms, plus one baseline method, to disambiguate among the candidate entities of each detected mention. Those systems are the same mentioned in (Barrena, Agirre and Soroa, 2014), which scored near the best system reported in accuracy at TAC-KBP 2013. The only difference is that this year we do not rank the NIL entity as we do in previous year.

The baseline consists in returning the most frequent entity (*MFE*) for each mention, according to the prior probability in the dictionary.

[1] http://nlp.stanford.edu/downloads/corenlp.shtml

[2] http://wiki.dbpedia.org/Ontology

[3] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/

[4] http://ixa2.si.ehu.es/ixa-pipes/

[5] APW_ENG_20101006.0253

First method, called The *iXanpei* and based on (Han and Sun, 2011), is a generative entity linking model that combines evidences from different probabilities to rank the candidates. It depends on the prior probability of the dictionary combined with textual context probability to rank candidates.

The second method, called *UKB* (Agirre and Soroa, 2009; Agirre, Lopez de Lacalle and Soroa, 2014)[6], is a graph based method for disambiguation which represents the Wikipedia pages and hyperlinks as a graph. This method ranks candidates taking into account the weights given by a Personalized PageRank algorithm.

Finally, the third method, called *RF*, is a combination of the previous two methods by means of a supervised classifier based on Random Forests.

## 5 Knowledge base Mapping and NIL clustering

Our systems return entities from the 2011 Wikipedia snapshot, and we need to link them to the TAC-KBP knowledge base[7]. If there is no direct match, we test whether there is any reference KB entity which redirects to the entity returned by the system, according to the 2011 version of Wikipedia. If so, we return the KB entity and if not we return NIL. All NIL mentions associated with the same Wikipedia 2011 entity are grouped in the same cluster. Finally, if the mention does not occur in the dictionary we also return NIL. In this case the clustering is very basic: all mentions having the same surface form are assigned the same cluster.

## 6 Classifing mention-entity pairs

Finally, we classify the disambiguated entities as *PER*, *ORG* or *GPE* taking into account the result of disambiguation system. If, according to the reference KB, the entity is of one of such types we just return it. However, if the type of the reference KB entity is UNK, or in case of NILs, we use the resource which maps entities to corresponding categories mentioned on 2. Note that those disambiguated entity mentions that can not be linked to this resource are discarded. This way we discard mentions like mentioned in section 3 because the resulting entity is not classified as *PER*, *GPE* or *ORG*. This way we deal with the mentions over generated by *Match-up*.

We also tried to assign the type given by the *Ixa-pipe-nerc* tagger, with unsatisfactory results.

## 7 Evaluation measures

This year organizers used several measures to evaluate different aspects of EDL task, such as name tagging, linking and clustering. The official measures were defined as clustering performance (CEAFmF1) and linking performance (WIKIF1). CEAFmF1 aligns system and gold mentions based on span offsets, but it does not require the entity type or the KB identifier to match. It then performs standard mention CEAF calculation. WIKIF1 computes the Entity Linking F-measure, considering both entity and type matching, but leaving aside mention detection. The organizers also used the DISCF1 measure, which evaluates mention detection and whether the mention is linked to an KB entity or to NIL. Finally, DEL task is evaluated using Bcubed+ F1.

## 8 Experimental results

For the EDL task we used two methods to detect mentions (*Match-up*, *Ixa-pipe-nerc*) and three disambiguation systems (*iXanpei*, *UKB*, *RF*), that is, a total of six detection-disambiguation combinations. Clustering and entity classification is the same in all systems. We sent five runs depending on the results obtained by the systems in the training dataset[8]. These are combinations of mention detection and disambiguation for the submitted five runs:

- EDL_Run1: *Match-up* & *UKB*
- EDL_Run2: *Match-up* & *iXanpei*
- EDL_Run3: *Ixa-pipe-nerc* & *UKB*
- EDL_Run4: *Ixa-pipe-nerc* & *iXanpei*
- EDL_Run5: *Ixa-pipe-nerc* & *RF*

Table 1 shows the performance for the submitted runs in the training dataset. The *Match-up* and

[6] `http://ixa2.si.ehu.es/ukb`

[7] The reference KB for TAC is a subset of a 2008 dump of Wikipedia.

[8] LDC2014E54, TAC 2014 KBP English Entity Discovery and Linking Training Data

| Systems | WIKIF1 | DISCF1 | LINKF1 | *CEAFmF1* |
|---|---|---|---|---|
| *Match-up* & *UKB* | **0.651** | 0.655 | **0.616** | **0.655** |
| *Match-up* & *iXanpei* | 0.617 | 0.646 | 0.600 | 0.652 |
| *Ixa-pipe-nerc* & *UKB* | 0.607 | 0.640 | 0.584 | 0.635 |
| *Ixa-pipe-nerc* & *iXanpei* | 0.577 | 0.643 | 0.578 | 0.633 |
| *Ixa-pipe-nerc* & *RF* | 0.617 | **0.656** | 0.602 | 0.652 |

Table 1: Entity Discovery and Linking WIKIF1, DISCF1, LINKF1 and CEAFmF1 measures for training dataset. Bold marks best performance for each measure.

| Run | WIKIF1 | DISCF1 | LINKF1 | *CEAFmF1* |
|---|---|---|---|---|
| EDL_Run1 (*Match-up* & *UKB*) | **0.613** | 0.632 | **0.598** | 0.618 |
| EDL_Run2 (*Match-up* & *iXanpei*) | 0.560 | 0.628 | 0.570 | 0.610 |
| EDL_Run3 (*Ixa-pipe-nerc* & *UKB*) | 0.603 | 0.636 | **0.598** | 0.618 |
| EDL_Run4 (*Ixa-pipe-nerc* & *iXanpei*) | 0.543 | 0.646 | 0.568 | 0.615 |
| EDL_Run5 (*Ixa-pipe-nerc* & *RF*) | 0.588 | **0.655** | **0.598** | **0.629** |
| **Best** | 0.678 | 0.749 | 0.673 | 0.730 |
| **Rank10** | 0.509 | 0.600 | 0.517 | 0.559 |

Table 2: Entity Discovery and Linking WIKIF1, DISCF1, LINKF1 and CEAFmF1 measures for our submitted runs compared to best and Rank 10 performance. Bold marks best performance for each measure.

*UKB* combination attains the best scores according to most evaluation measures. We also consider sending the probabilistic model with the same mention detection system. In order to test both *Match-up* and *Ixa-pipe-nerc* we choose to send similar runs, that is combined with the same disambiguation algorithms. Done this, we send a fifth run to test the RF classifier that scored similar to the best runs.

Table 2 shows the results of in the test dataset. Our best score according to CEAFmF1 was obtained when combining *Ixa-pipe-nerc* for mention detection and *RF* classifier for disambiguation. This system is among the top 10 systems[9], but 10 points below the best system. Our best score according to WIKIF1 was scored combining *Match-up* and *UKB*. This system is again among the top 10 systems, 8 points better than the system ranked 10th and circa 7 points below the best system. The results obtained according to DISCF1 shows that *Ixa-pipe-nerc* does slightly better than *Match-up* in mention detection, but that both mention detection algorithms need to improve. Note that *Ixa-pipe-nerc* was trained in CoNLL 2003 dataset, which followed different guidelines than those followed by the organizers when annotating the named entities from the test dataset. We suspect that the performance of *Ixa-pipe-nerc* was adversely affected by this mismatch.

Regarding the DEL, we sent 4 runs, one for each disambiguation method described above (*RF*, *UKB*, *iXanpei*) plus the baseline method. Table 3 shows the Bcubed+ and accuracy results for the DEL track. Note that accuracy does not consider NIL clustering. We obtained close to top results on accuracy and Bcubed+ F1 for in-KB instances, reaching 0.772 Bcubed+ F1 score with *UKB* system. Comparing Bcubed+ and accuracy results, we see that clustering algorithm needs to improve in order to keep the accuracy based results. All in all, our system attained 5 points above the median score and 7 points below the best system in All (inKB + NIL) queries.

## 9 Conclusions and future work

Due to the yield loss in mention detection shown at DISCF1 scores, our systems have been also penalized in CEAFmF1 measure. Even so, we have seen that the performance of the mention detection based on Wikipedia anchors *Match-up* is very similar to *Ixa-pipe-nerc*. In addition, we have seen that the strategy of discarding mentions that can not be linked to entity-type pairs resource is valid.

[9] EDL task gather the results from 20 teams and a total of 74 runs.

| Run | All | inKB | Nil |
|---|---|---|---|
| DEL_Run1 (*RF*) | 0.747 (*0.804*) | 0.771 (*0.811*) | 0.719 (*0.795*) |
| DEL_Run2 (*UKB*) | **0.752** (*0.820*) | **0.772** (*0.812*) | **0.729** (*0.828*) |
| DEL_Run3 (*iXanpei*) | 0.701 (*0.763*) | 0.696 (*0.745*) | 0.708 (*0.784*) |
| DEL_Run1 (*MFE baseline*) | 0.676 (*0.738*) | 0.628 (*0.677*) | 0.731 (*0.810*) |
| **Best** | 0.821 (*0.868*) | 0.796 (*0.827*) | 0.855 (- - - -) |
| **Median** | 0.698 (*0.769*) | 0.648 (*0.669*) | 0.767 (- - - -) |

Table 3: Diagnostic Entity Linking Bcubed+ F1 and accuracy measures (in brackets) for our submitted runs compared to best and median performance. Bold marks best performance for each measure.

In the other hand, our NED systems have done a good work. Both results in WIKIF1 for EDL and Bcubed+ F1 (in KB) for DEL shows that our systems obtain very good results and they are close to the best performance in both tracks.

Considering the different guidelines and datasets used for training the NERC tagger, in the future we want to re-train the *Ixa-pipe-nerc* tagger in a dataset such as the additional ERE annotations of Discussion Forum documents[10] available at LDC. We expect a significant boost in mention detection performance in doing do. Besides, we do not discard the possibility of combine both systems, introducing *Match-up* detection results as a feature for *Ixa-pipe-nerc*.

We also plan to implement more sophisticated clustering algorithms in order to improve clustering results, using acronym expansion or name similarity techniques.

## Acknowledgements

The research leading to these results was carried out as part of the READERS project (`http://nlp.uned.es/readers-project/`) funded by European Communitys in the framework of ERA-NET CHIST-ERA. The work has been also carried out as part of the by QTLeap project (FP7-ICT-2013.4.1-610516). Ander Barrena enjoys an scholarship from the Basque Government.

## References

Eneko Agirre and Aitor Soroa. 2009. *Personalizing PageRank for Word Sense Disambiguation* Proceedings of 14th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece.

Eneko Agirre, Oier Lopez de Lacalle and Aitor Soroa. 2014. *Random Walks for Knowledge-Based Word Sense Disambiguation* Computational Linguistics, volume 40, number 1, pages 57-88.

Xianpei Han and Lee Sun. 2011. *A generative entity-mention model for linking entities with knowledge base*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. Stroudsburg, PA, USA.

Ander Barrena, Eneko Agirre and Aitor Soroa. 2014. *UBC Entity Linking at TAC-KBP 2013: random forests for high accuracy* Text Analysis Conference, Knowledge Base Population 2013 Gaithersburg, Maryland, USA.

Rodrigo Agerri, Josu Bermudez and German Rigau. 2014. *IXA pipeline: Efficient and Ready to Use Multilingual NLP tools* Proceedings of the 9th Language Resources and Evaluation Conference LREC2014 Reykjavik, Iceland.

[10] LDC2014E31, DEFT ERE English Discussion Forum Annotation V3