# Wikipedia Search as Effective Entity Linking Algorithm

Milan Dojchinovski<sup>1,2</sup>, Tomáš Kliegr<sup>1</sup>, Ivo Lašek<sup>1,2</sup> and Ondřej Zamazal<sup>1</sup>

Department of Information and Knowledge Engineering Faculty of Informatics and Statistics University of Economics, Prague, Czech Republic

name.surname@vse.cz <sup>2</sup> Web Engineering Group Faculty of Information Technology Czech Technical University in Prague, Czech Republic name.surname@fit.cvut.cz

#### Abstract

This paper reports on the participation of the LKD team in the English entity linking task at the TAC KBP 2013. We evaluated various modifications and combinations of the Most-Frequent-Sense (MFS) based linking, the Entity Co-occurrence based linking (ECC), and the Explicit Semantic Analysis (ESA) based linking. We employed two our Wikipediabased NER systems, the Entityclassifier.eu and the SemiTags. Additionally, two Lucenebased entity linking systems were developed. For the competition we submitted 9 submissions in total, from which 5 used the textual context of the entities, and 4 submissions did not. Surprisingly, the MFS method based on the Wikipedia Search has proved to be the most effective approach – it achieved the best  $0.555 \text{ B}^{3+} \text{ F1}$  score from all our submissions and it achieved high 0.677 B<sup>3+</sup> F1 score for Geo-Political (GPE) entities. In addition, the ESA based method achieved best 0.483 B<sup>3+</sup> F1 for Organization (ORG) entities.

# 1 Introduction

In the last decade the number of Named Entity Recognition (NER) systems which recognize, classify and link entities in text with entities in other knowledge bases is constantly increasing. One of the key tasks of a NER system is the *entity linking* task. Its ultimate goal is to enable linkage of text corpora (not structured information) with other knowledge bases (structured information).

In order to enable linking of entities in text with other knowledge bases, in our previous work we have developed the Entityclassifier.eu <sup>1</sup>(Dojchinovski and Kliegr, 2013) and SemiTags<sup>2</sup> (Lašek and Vojtáš, 2013) NER systems. Both systems can spot entities in text, link them with entities in Wikipedia (resp. DBpedia) and finally, classify them with concepts from the DBpedia Ontology.<sup>3</sup> While the Entityclassifier.eu system performs entity linking based on the Most-Frequent-Sense (MFS) method and does not use the text context around the entities, the SemiTags utilizes more advanced entity linking method which is based on Entity Cooccurrence (ECC) in Wikipedia and uses the entity text context.

In this paper, we report on the evaluation of the entity linking methods of the Entityclassifier.eu and Semitags NER systems on the TAC KBP 2013 Entity Linking task.<sup>4</sup> For the task we also developed and evaluated additional entity linking method which is based on the Explicit Semantic Analysis (ESA) approach (Gabrilovich and Markovitch, 2007). We also evaluated two additional variants of the MFS method used by the Entityclassifier.eu system.

The reminder of the paper is structured as follows. Section 2 describes how the provided TAC KBP knowledge base was prepared and linked with our Wikipedia (resp. DBpedia) knowledge base. Section 3 describes our entity linking methodology, evaluated entity linking methods and provides de-

<sup>&</sup>lt;sup>1</sup>http://entityclassifier.eu/

<sup>&</sup>lt;sup>2</sup>http://ner.vse.cz/SemiTags/

<sup>&</sup>lt;sup>3</sup>http://dbpedia.org/Ontology

<sup>&</sup>lt;sup>4</sup>http://www.nist.gov/tac/2013/KBP/ EntityLinking/

scription for each of the 9 submissions. Section 4 presents and discusses our achieved results. Finally, Section 5 concludes the paper.

# 2 Task Description and Data Preparation

## 2.1 The Entity Linking Task

The entity linked task, as described by the challenge organizers<sup>5</sup> is described as task of linking entity mentions in a document corpora to entities in a reference KB. If the entity is not already in the reference KB, a new entity node should be added to the KB. Each participation team was given a set of 2190 queries consisting of a *queryID*, *docID*, *name* (name mention of an entity) and *beg* and *end* entity offsets in the document.

Further, the system performing the entity linking task had to output the results providing information about the queryID,  $KB\_link$  (or NIL entity identifier, if the entity was not present in the KB) and a *confidence value*.

## 2.2 Data Preparation

Since the TAC KBP reference knowledge base uses custom identifiers of the entities (e.g. E0522900), and our systems identify the entities with DBpedia URIs, it was necessary to perform mapping of these identifiers.

In the TAC KBP knowledge base each entity entry provides information about the custom identifier of the entity, and the path URL segment of a Wikipedia article describing the entity (e.g. entity with URL *Sam\_Butler* and id *E*0522900). Since DBpedia also derives its URIs from the URIs of the Wikipedia articles, we used the URLs of the Wikipedia articles describing the entities to map them to DBpedia. For example, the entity in the TAC KBP KB with identifier *E*0522900 was mapped to DBpedia URI http://dbpedia.org/resource/ Sam\_Butler. This way we could relate the DBpedia URI identifiers of our systems with the entity identifiers in the TAC reference knowledge base.

# 3 Methodology

In our Entity Linking approach we have developed and used three different independent methods. Two of the methods, the *MFS* method and the *ECC* method, are already used for entity linking in the Entityclassifier.eu and SemiTags NER systems respectively. A third, novel method based on the *ESA* approach was additionally developed for the entity linking challenge.

All three methods follow the three-steps approach defined as follows. First, it applies *candidate selec-tion*, where a set of entity candidates are retrieved for the given entity. Second, it performs the *disam-biguation*, where one candidate from the candidates list is selected as the correct one. Finally, selected entity candidate is linked, i.e. a reference in the TAC KBP knowledge base is identified. If the entity is not found in the KB, then, a new entity NIL node is added.

Bellow we describe each used method for entity linking followed by a description of each submission.

# 3.1 Most Frequent Sense Method

The MFS method is a context independent method which does not use the context text around the entity, but it uses only the entity name when performing the linking. In this approach the entity is linked with the most-frequent-sense entity found in the reference knowledge base. To realize the MFS approach we used the available English Wikipedia Search API and we also used a specialized Lucene index<sup>6</sup> which extends the Apache Lucene search API. It primarily ranks pages based on number of backlinks and the Wikipedia articles' titles. Note that the Wikipedia Search API is build on top of the Lucene index and it offers some more functionalities. This approach was considered as a baseline.

## 3.2 Entity Co-occurrence Method

By ECC method, we aim at capturing relations between entities rather than their textual representation. An example of a similar structural representation is described in (Milne and Witten, 2008), where each entity (represented as a Wikipedia article) is characterized by a structure of incoming links instead of its textual description.

Contrary to the approach presented in (Milne and Witten, 2008) our ECC based model does not

<sup>&</sup>lt;sup>5</sup>http://www.nist.gov/tac/2013/KBP/ EntityLinking/

<sup>&</sup>lt;sup>6</sup>http://www.mediawiki.org/wiki/ Extension:Lucene-search

compare similarities of individual entities. We are searching for the best combination of candidates (possible meanings) for individual surface forms in an analysed text, where individual paragraphs represent the context.

For example, let us consider the following sentence: *Michael Bloomberg was the mayor of New York*. Simple observation shows that the entity Michael Bloomberg (former mayor of New York) co-occurs in the same paragraph in our knowledge base together with the correct entity New York City in United States much more often (88 times) than with the New York in England (0 times).

Because generating all candidate combinations is a very demanding task, we developed a heuristic that quantifies an impact of co-occurrences in the same paragraph.

We construct an incidence matrix I of the size  $|C| \times |C|$ , where C is the set of candidate entities (possible meanings of the identified name). The weights in the matrix are the co-occurrence measures.

In our case we measure number of paragraphs where the two candidates occur in the same paragraph in our knowledge base (Wikipedia).

Then we compute a score  $e_{i,s}$  for each candidate as a sum of a line of the matrix representing the candidate (Equation 1).

$$e_{i,s} = \sum_{j=1}^{|C|} e_{i,j}$$
 (1)

**Indexing Wikipedia for Contextual Disambiguation.** In order to parse and index Wikipedia dump files, we implemented our own parser using gwtwiki parser.<sup>7</sup> We extract individual paragraphs of Wikipedia articles and identify contained links together with their anchor texts. The resulting indexes are stored in Redis<sup>8</sup> in-memory store for faster retrieval.

#### 3.3 Explicit Semantic Analysis Method

In the ESA method (Gabrilovich and Markovitch, 2007), the input text T is represented as a TF-IDF term vector. For each word  $w_i$  in the input text the

method uses an inverted index to retrieve Wikipedia articles  $c_1, \ldots, c_n$  containing  $w_i$ . The semantic relatedness of the word  $w_i$  with concept  $c_j$  is computed such that the strength of association between  $w_i$  and  $c_j$  is multiplied with the TF-IDF weight of  $w_i$ in T. The relatedness score for any two documents is determined by computing the cosine similarity between the vectors of document-concept semantic relatedness.

ESA has a number of a follow-up papers describing particularly its applications in various areas of information retrieval, including cross-language information retrieval (cf. (Gottron et al., 2011) for an overview). The use of ESA for disambiguation of a surface form to a Wikipedia URL was proposed in (Fernandez et al., 2011).

#### 3.4 Submissions Description

For the TAC KBP 2013 Entity Linking task we have submitted 9 runs based on different variations of the MFS, ECC and ESA methods. Bellow we provide detailed description of each individual submission. The first four runs rely on the MFS linking approach, the fifth run rely on the entity ECC linking approach, the sixth on the ESA linking approach, the seventh is a merged submission of the ECC and ESA linking approaches, the eight is a combination of the MFS and ESA linking approaches, and finally, the ninth is a merged submission of the MFS, ECC and ESA linking approaches.

**Run #1.** This run relies on the MFS approach to perform the linking. In this run each entity mention was considered as an entity, so the entity spotting step was not performed. To realize the MFS approach we used the English Wikipedia Search API. For the entity candidate selection step we run a Wikipedia search for the most five relevant articles describing the entity in question. The article with the highest rank in the result list was considered as the correct entity. Finally, this entity was linked with an entity in the knowledge base. If it was not found, then, a new entity NIL node was created.

This run did access the Web during the processing and it did not use the entity context text (wiki\_text element text) neither it uses the entity offsets.

**Run #2.** This run also relies on the MFS linking approach. Compared to the previous run, in this run

<sup>&</sup>lt;sup>7</sup>Gwtwiki parser - project homepage: http://code.google.com/p/gwtwiki/

<sup>&</sup>lt;sup>8</sup>Redis key-value store homepage: http://redis.io/

we used the Entitiyclassifier.eu <sup>9</sup> NER system to perform the linking. In this run each entity mention was submitted to the NER system. The system decides whether the string represents an entity. In a positive scenario the system disambiguates the entity by running a Wikipedia search on the API for the local English Wikipedia mirror. Similarly like in the previous run, the highest ranked result is considered as the correct entity. The NER system finally returns a DBpedia resource URI which describes the entity. This entity URI is then linked to the entity in the reference knowledge base. If it was not found, then, a new entity NIL node was created.

Note that this run did not access the Web during the processing and it does not use the entity context text (wiki\_text element text), neither it uses the entity offsets.

**Run #3.** This run, same as the previous two runs, relies on the MFS linking approach. This run uses a local Lucene index created for an English Wikipedia snapshot, as of 18/9/2012. Note that this Lucene index is also used by the Wikipedia Search API. Each entity mention was searched in the index and the first returned result was considered as the correct entity. Unlike the previous runs, this run provides direct reproducibility, as it involves no third-party hosted software or data. Since this run uses a local Lucene index, it did not access the Web during the processing, neither it uses the entity context text or the entity offsets.

**Run #4.** This is a merged submission of the previous three MFS linking approaches. In the case that we had different entity links for the same query we experienced that merging the results produced conflicts very often. These conflicts were resolved based on the performance of the individual methods on the TAC 2012 corpus. Thus, run #2 had the highest priority followed by the run #1, and finally, run #3.

Since this run merges results from two submissions that do not access the Web and one that accesses the Web, it can be considered that this run accesses the Web too. On the other side, this submission does not use the entity context text. Run #5. This run relies on the ECC method used for entity linking. In this run we used the SemiTags NER system to perform the linking. See Section 3.2 for detailed description of the ECC method and the SemiTags NER system. For each query we submitted 800 characters long text to the NER system. Submitted text consists of the entity and 400 characters preceding and following the entity. To create this context text we used the text from the wiki\_text element. The NER system recognizes entities in the text, links those entities with DBpedia resource URIs and returns the results. Next, we check whether the entity in question is recognized as an entity and whether a DBpedia link is provided. In a positive scenario, the DBpedia URI is mapped to an entity in the reference knowledge base. If the entity was not present in the KB, a new entity NIL node was added. This run uses the entity context text (wiki\_text element) but does not access the Web.

Run #6. This run relies on the ESA entity linking method. In this method each entity mention is considered as an entity. For each query, top five results returned by Wikipedia Search API are used as entity candidates. Next, the first paragraph of each of these candidate articles is retrieved. Finally, the ESA<sup>10</sup> method is used to compute the similarity of the first paragraph of each entity candidate with the entity context text. The entity context text is a 800 characters long text constructed from the wiki\_text element. Same as in the previous submission, it consists the entity and 400 characters preceding and following the entity. After computing the similarity between each first entity paragraph and the entity context text, the entity (first Wikipedia article paragraph) with the highest similarity score is considered as correct and it was linked with the entity in the reference KB.

This run does not access the Web during the evaluation.

**Run #7.** This is a merged submission of the ESA and the ECC linking methods which uses the wiki\_text. A higher priority was given to the ECC method when resolving the conflicts. Since none of the two submissions access the Web, it can be consistently considered that this run does not access the

<sup>&</sup>lt;sup>9</sup>Note that for the run #2 we used a local instance of the NER system so we consider this run as run that does not access the Web. The API documentation of the Entityclassifier.eu NER system is publicly available at: http:// entityclassifier.eu/thd/docs/

<sup>&</sup>lt;sup>10</sup>Employed ESA implementation is available at: http://ticcky.github.io/esalib/

Web either.

Run #8. This run combines the MFS and the ESA entity linking approaches. In this run for each query, we used the Wikipedia Search API to retrieve a list of entity candidates. We used only the first result as a candidate (cf. run #1). Next, the ESA method was used to compute the similarity of the first paragraph of the Wikipedia article describing the entity candidate with the entity context text. As the context of the entity 400 characters preceding and following the query (800 altogether) was used. In addition, threshold of 0.15 was set for the ESA linking method. If the similarity score was higher or equal to 0.15, then the entity was considered as correct and the entity was linked with the entity in the reference KB. Otherwise, it was considered as incorrect and a new entity NIL node was added.

This run accesses the Web during the evaluation and it is using the entity context text (wiki\_text element).

**Run #9.** This run is a merged submission of four individual submissions. The conflicts were resolved by assigning priority to each individual submission. The highest priority was given to the run #2 (MFS with Entityclassifier.eu NER), followed by run #5 (SemiTags NER), run #6 (ESA) and run #1 (MFS baseline). Since some runs access the Web and also use the entity context (wiki\_text), it can be considered that this submission accesses the Web and uses the entity context too.

In all the runs the basic "exact name" NIL clustering technique was only used.

### 4 Evaluation

### 4.1 Metrics

In the TAC 2013 KBP Entity Linking task the systems were evaluated using three scoring metrics. The micro-average ( $\mu AVG$ ) (McNamee et al., 2010), the B-cubed cluster scoring ( $B^3$ ) (Bagga and Baldwin, 1998) and the B-cubed+ modification ( $B^{3+}$ ) (Ji et al., 2011). The main difference between the B-cubed and B-cubed+ scoring metrics is that for non-NIL queries, B-cubed+ not only consider the quality of clustering as in B-cubed, but also measure the accuracy of linking these queries to correct KB entries.

### 4.2 Results

We report all three scoring metrics for each of our submissions, as well we report on the performance of our methods for the focus queries (how well our methods perform for entities of type person, organization, etc.). Note that we also report on the precision and recall for each metric and the median value for the all participating teams.

In Table 1 we provide the overall performance achieved of each individual run. The results show that in overall, the MFS linking method (cf. run #1) performed the best, achieving 0.707 B-cubed F1 score, 0.555 B-cubed+ F1 score and highest Bcubed+ precision score 0.653. The highest B-cubed precision score was achieved by the ECC linking method with score of 0.912.

Compared with the reported median B-cubed+ F1 score at the challenge, the MFS based linking submissions (#1 - #4) achieved similar score for all the queries (see Table 2). However, for the focus queries targeting GPE entities the MFS submission achieved significantly better B-cubed+ F1 score 0.677 compared with the reported median at the challenge 0.552. A better B-cubed+ F1 compared to the median was also achieved for the focus queries targeting entities in discussion fora (0.539 compared to median 0.488).

Surprisingly, the MFS method based submissions, which do not use the entity context (wiki\_element text) showed better results compared with the ECC and ESA method based submissions which use the entity context text.

The queries in the Entity Linking task were targeting entities of three different types. Entities of type Person (PER), Organization (ORG) and Geo-Political entities (GPE). For the task each participating team received 2190 queries in total where 686 queries were targeting entities of type person, 701 targeting organizations and 803 targeting geo-political entities. Tables 3–5 summarize the achieved results related to these focus queries for each individual submission.

Id	$\mu AVG$	<b>B</b> <sup>3</sup> <b>P</b>	<b>B<sup>3</sup> R</b>	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	<u>0.737</u>	0.870	<u>0.596</u>	<u>0.707</u>	0.653	<u>0.483</u>	0.555
run #2	0.686	0.821	0.515	0.633	0.568	0.387	0.461
run #3	0.727	0.844	0.593	0.696	0.632	0.471	0.540
run #4	0.733	0.835	0.570	0.678	0.622	0.461	0.530
run #5	0.611	<u>0.912</u>	0.428	0.582	0.558	0.292	0.383
run #6	0.625	0.896	0.500	0.642	0.562	0.358	0.437
run #7	0.658	0.901	0.499	0.642	0.604	0.373	0.462
run #8	0.717	0.887	0.546	0.676	0.640	0.433	0.517
run #9	0.704	0.850	0.580	0.690	0.610	0.461	0.525

Table 1: Overall performance of all the runs.

Focus Queries	Highest B <sup>3+</sup> F1	Median B <sup>3+</sup> F1	Our highest B <sup>3+</sup> F1
All	0.746	0.574	0.555
in KB	0.722	0.554	0.595
not in KB	0.777	0.566	0.601
NW - Newswire docs	0.829	0.645	0.586
WB - Web docs	0.678	0.525	0.484
DF - Discussion Fora docs	0.662	0.488	0.539
PER	0.778	0.627	0.501
ORG	0.737	0.542	0.483
GPE	0.746	0.552	0.677

Table 2: Comparison of the highest and median  $B^{3+}F1$  scores achieved at the challenge with the highest score achieved by our submissions.

ID	$\mu AVG$	В <sup>3</sup> Р	B <sup>3</sup> R	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	0.756	0.966	0.733	0.833	0.739	0.624	0.677
run #2	0.663	0.930	0.581	0.716	0.630	0.461	0.532
run #3	0.737	0.955	0.724	0.824	0.721	0.605	0.658
run #4	0.757	0.942	0.684	0.792	0.723	0.595	0.653
run #5	0.513	0.949	0.412	0.575	0.491	0.272	0.350
run #6	0.553	0.963	0.508	0.665	0.539	0.359	0.431
run #7	0.624	0.941	0.536	0.683	0.596	0.424	0.495
run #8	0.704	<u>0.968</u>	0.649	0.777	0.687	0.531	0.599
run #9	0.737	0.932	0.681	0.787	0.698	0.590	0.640

Political) entities.

ID	$\mu AVG$	B <sup>3</sup> P	<b>B<sup>3</sup> R</b>	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	0.691	0.856	0.559	0.676	0.604	0.428	0.501
run #2	0.636	0.824	0.511	0.630	0.530	0.357	0.427
run #3	0.669	0.806	0.558	0.660	0.563	0.409	0.474
run #4	0.691	0.837	0.553	0.666	0.591	0.420	0.491
run #5	0.570	<u>0.895</u>	0.425	0.577	0.511	0.269	0.352
run #6	0.586	0.869	0.487	0.624	0.512	0.324	0.397
run #7	0.620	0.883	0.479	0.621	0.561	0.330	0.416
run #8	0.665	0.874	0.505	0.641	0.584	0.380	0.461
run #9	0.672	0.852	0.555	0.672	0.589	0.417	0.489

Table 5: Results for queries targeting PER (Person) entities.

ID	$\mu AVG$	<b>B</b> <sup>3</sup> <b>P</b>	B <sup>3</sup> R	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	0.762	0.772	0.477	0.589	0.604	0.374	0.462
run #2	0.762	0.693	0.443	0.541	0.533	0.333	0.410
run #3	0.773	0.754	0.476	0.583	0.597	0.378	0.463
run #4	0.746	0.711	0.457	0.557	0.536	0.348	0.422
run #5	0.763	<u>0.886</u>	0.447	0.594	0.682	0.336	0.450
run #6	0.745	0.846	0.503	0.631	0.636	0.390	<u>0.483</u>
run #7	0.736	0.874	0.476	0.616	0.656	0.358	0.464
run #8	0.783	0.807	0.468	0.592	0.640	0.373	0.472
run #9	0.698	0.755	0.488	0.593	0.529	0.355	0.425

Table 3: Results for queries targeting GPE (Geographical- Table 4: Results for the queries targeting ORG (Organization) entities.

ID	$\mu AVG$	<b>B<sup>3</sup> P</b>	<b>B<sup>3</sup> R</b>	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	0.668	0.934	0.601	0.732	0.635	0.469	0.539
run #2	0.578	0.917	0.483	0.632	0.537	0.320	0.401
run #3	0.658	0.884	0.596	0.712	0.611	0.452	0.519
run #4	0.662	0.925	0.562	0.699	0.624	0.432	0.511
run #5	0.522	<u>0.953</u>	0.376	0.539	0.499	0.220	0.305
run #6	0.539	0.937	0.460	0.617	0.503	0.303	0.378
run #7	0.593	0.939	0.460	0.617	0.564	0.325	0.412
run #8	0.603	0.941	0.526	0.674	0.572	0.378	0.456
run #9	0.640	0.919	0.560	0.696	0.600	0.427	0.499

Table 6: Results for queries from the English discussion forum documents selected from the BOLT Phase 1 forum data.

ID	$\mu AVG$	B <sup>3</sup> P	<b>B<sup>3</sup> R</b>	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	<u>0.714</u>	0.857	0.500	0.631	0.630	0.392	<u>0.484</u>
run #2	0.706	0.819	0.455	0.585	0.576	0.336	0.424
run #3	0.703	0.832	0.496	0.622	0.605	0.377	0.465
run #4	0.706	0.824	0.481	0.607	0.589	0.365	0.451
run #5	0.665	<u>0.923</u>	0.432	0.588	0.623	0.304	0.409
run #6	0.647	0.893	0.472	0.617	0.594	0.338	0.431
run #7	0.653	0.902	0.472	0.620	0.611	0.337	0.434
run #8	0.729	0.888	0.467	0.612	0.661	0.374	0.478
run #9	0.650	0.838	0.494	0.622	0.558	0.355	0.434

from various GALE web collections.

ID	$\mu AVG$	<b>B<sup>3</sup> P</b>	B <sup>3</sup> R	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	0.903	0.813	0.523	0.636	0.734	0.487	0.586
run #2	<u>0.953</u>	0.742	0.478	0.582	0.710	0.461	0.559
run #3	0.910	0.788	0.521	0.627	0.715	0.489	0.581
run #4	0.885	0.756	0.477	0.585	0.663	0.433	0.524
run #5	0.929	0.888	0.436	0.585	0.832	0.404	0.544
run #6	0.835	0.852	0.495	0.626	0.717	0.421	0.531
run #7	0.797	0.878	0.434	0.581	0.712	0.345	0.465
run #8	0.942	0.832	0.507	0.630	0.788	0.486	0.601
run #9	0.749	0.782	0.453	0.574	0.580	0.345	0.432

Table 9: Results for queries targeting entities which are not present in the KB.

The achieved results for the GPE and PER focus queries (see Table 3 and 5) show that our MFS linking method (cf. run #1) performed best achieving Bcubed+ F1 score 0.677 for GPE and 0.501 for PER focus queries. However, the MFS performed worse for the ORG (see Table 4) focus queries. One of the reasons for such performance of the MFS method can be that the GPE and PER entities are less ambiguous compared to the ORG entities.

The results also show that the ESA linking method (cf. run #6) achieved highest B-cubed+ F1 score 0.483 (see Table 4) for the ORG focus queries.

In the Entity Linking task the queries were targeting entities in three collections of documents. The Newswire collection (NW) consisting of documents from the English Gigaword Fifth Edition, the Web collection (WB) consisting of documents from various GALE web collections, and the Discussion Fora collection (DF) consisting of documents selected from the BOLT Phase 1 discussion forums source data. Tables 6-8 summarize the results for queries targeting entities in these three collections.

In the three collections the ECC method, which takes into account the context of the entities, achieved highest B-cubed precision. However, in overall the MFS linking based submissions achieved

ID	$\mu AVG$	В <sup>3</sup> Р	B <sup>3</sup> R	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	0.788	0.833	0.622	0.712	0.672	0.519	0.586
run #2	0.748	0.762	0.554	0.641	0.585	0.446	0.506
run #3	0.779	0.822	0.620	0.707	0.653	0.511	0.573
run #4	0.786	0.782	0.603	0.681	0.631	0.509	0.563
run #5	0.651	<u>0.883</u>	0.459	0.604	0.576	0.333	0.422
run #6	0.672	0.872	0.534	0.662	0.588	0.399	0.475
run #7	0.701	0.877	0.532	0.662	0.628	0.415	0.500
run #8	0.785	0.853	0.582	0.692	0.676	0.485	0.565
run #9	0.761	0.811	0.619	0.702	0.632	0.514	0.567

Table 7: Results for the queries from the Web documents Table 8: Results for queries from the Newswire documents from the English Gigaword Fifth Edition.

ID	$\mu AVG$	<b>B<sup>3</sup> P</b>	B <sup>3</sup> R	B <sup>3</sup> F1	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F1
run #1	0.597	0.918	0.659	0.767	0.585	0.479	0.527
run #2	0.458	0.888	0.546	0.676	0.447	0.325	0.376
run #3	0.572	0.892	0.654	0.755	0.561	0.456	0.503
run #4	0.604	0.902	0.650	0.755	0.587	0.486	0.532
run #5	0.340	0.933	0.420	0.579	0.325	0.196	0.244
run #6	0.445	0.934	0.504	0.655	0.429	0.304	0.356
run #7	0.540	0.921	0.554	0.692	0.512	0.398	0.448
run #8	0.525	0.934	0.579	0.715	0.514	0.388	0.442
run #9	<u>0.666</u>	0.908	0.688	0.783	0.636	0.559	<u>0.595</u>

Table 10: Results for queries targeting entities which are present in the KB.

the highest B-cubed+ F1 scores in all three submissions.

Finally, we also evaluated the performance of each individual submission for queries targeting entities which are present and not present in the reference KB. For the focus queries targeting entities not in the KB, the submission #8, which combines the MFS and ESA methods achieved highest Bcubed+ F1 score 0.601. The submission #9 which is a merged submission of MFS, ECC and ESA methods, achieved highest B-cubed+ F1 score 0.595 for the focus queries targeting entities in the KB. The results are reported in Tables 9 and 10.

### 4.3 Findings

We hereby summarize the main findings from the evaluation:

- The MFS based linking method in overall achieved the best results. It achieved best Bcubed+ F1 score from all submitted runs.
- The context based ECC linking method achieved high B-cubed precision in general, as well as for the GPE, ORG and PER focus queries.
- The context based ESA linking method

achieved best B-cubed+ F1 score for ORG focus queries.

- Submissions that merge results from the MFS, ECC and ESA methods achieved best Bcubed+ F1 score for focus queries that are targeting entities present in the reference knowledge base.
- For focus queries targeting entities in discussion fora documents, web documents, or news documents, the MFS method achieved the best B-cubed+ F1 score.

# 5 Conclusion

The performance of the entity linking sub-task of the available NER systems is of a significant importance. In this paper we report on performance of three entity linking methods which use and do not use the context text around the entities. The MFS based linking method which performs linking with the most-frequent-sense entity in the KB. The ECC based linking methods which takes into account the entities which appear in the surrounding context. And the ESA based linking which relies on the explicit semantic analysis method which estimates relatedness between two text documents.

According to the results from the challenge the MFS method, which is based on the Wikipedia Search, was the most effective entity linking approach.

Since the MFS and ECC based linking methods are used for entity disambiguation in the Entityclassifier.eu and SemiTags NER systems respectively, the results from this evaluation will help to better understand in which situations and domains the systems are performing bad and use this information to improve theirs performance.

## Acknowledgments

This research was supported by the European Union's 7th Framework Programme via the LinkedTV project (FP7-287911) and the LOD2 project (FP7-257943).

### References

- Dojchinovski Milan and Tomáš Kliegr. 2013 Entityclassifier.eu: Real-time Classification of Entities in Text with Wikipedia, In *Proceedings of the ECMLPKDD Conference* (ECMLPKDD13), Prague, Czech Republic.
- Evgeniy Gabrilovich and Markovitch Shaul. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis, In *Proceedings of the 20th international joint conference on Artifical intelligence* (IJCAI'07). Hyderabad, India.
- Gottron, Thomas and Anderka, Maik and Stein, Benno. 2011. Insights into explicit semantic analysis, In *Proceedings of the 20th ACM international conference on Information and knowledge management* (CIKM'11). Glasgow, Scotland, UK.
- Fernandez, Izaskun and Alegria, Iñaki and Ezeiza, Nerea. 2011. Semantic relatedness for named entity disambiguation using a small wikipedia, In *Proceedings of the 14th international conference on Text, speech and dialogue* (TSD'11). Pilsen, Czech Republic.
- Heng Ji, Ralph Grishman, and Hoa Dang. 2011. Overview of the TAC2011 knowledge base population track, In *TAC 2011 Proceedings Papers* (TAC2011).
- Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains, In *LREC Workshop on Linguistics Coreference*, pages 563-566.
- Paul McNamee, Hoa Trang Dang, Heather Simpson, Patrick Schone, and Stephanie Strassel. 2010. An evaluation of technologies for knowledge base population, In *Proceedings of the Seventh International LREC Conference*. (LREC'10) Valletta, Malta.
- David Milne and Ian H. Witten. 2008 Learning to link with Wikipedia, In *Proceeding of the 17th ACM conference on Information and knowledge management* (CIKM 08), New York, NY, USA.
- Ivo Lašek and Peter Vojtáš 2013 Various Approaches to Text Representation for Named Entity Disambiguation, In *International Journal of Web Information Systems* 9 (3), art. no. 17094595, pp. 242–259, Emerald Group Publishing Limited.