# Evaluation of Named Entity Recognition Tools on Microposts

Štefan Dlugolinský
Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava
Slovak Republic
{stefan.dlugolinsky@savba.sk}

Marek Ciglan
Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava
Slovak Republic
{marek.ciglan@savba.sk}

Michal Laclavík
Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava
Slovak Republic
{michal.laclavik@savba.sk}

*Abstract*—**In this paper we evaluate eight well-known Information Extraction (IE) tools on a task of Named Entity Recognition (NER) in microposts. We have chosen six NLP tools and two Wikipedia concept extractors for the evaluation. Our intent was to see how these tools would perform on relatively short texts of microposts. Evaluation dataset has been adopted from the MSM 2013 IE Challenge. This dataset contained manually annotated microposts with classification restricted to four entity types: PER, LOC, ORG and MISC.**

## I. Introduction

There are several information extraction (IE) tools available which perform very well on relatively long text documents. This paper provides an evaluation of such tools on a dataset of microposts. Similar work has been done in [1], where the authors evaluated news-trained Stanford NER on tweets. They demonstrated that news-trained NER classifiers rely heavily on capitalization, which is unreliable in tweets. In [2], the authors compared the performance of their proprietary NER classifier on a CoNLL dataset and a handmade Twitter dataset (1684 postings). The authors observed that although the NER classifier performed very well on a CoNLL dataset ($F_1$ over 83%), the performance on the Twitter dataset was extremely poor ($F_1$ below 40%). Another related evaluations have been done in [3] and [4]. In [4] authors evaluated semantic annotation platforms according to their annotation methods. They have examined two primary approaches: machine learning and pattern-based. The first one performed more effectively than the second one. Authors of [3] evaluated several NER systems on a manually built corpus. The corpus was very small (579 words in 13 paragraphs; in average 100 words per paragraph), but they have focused on the quality of its content to cover various typographic, lexical, semantic and heuristic features.

## II. Tools Evaluated

We have chosen six well-known NLP tools with the ability to recognize named entities in the text and two Wikipedia concept extractors. Our intent was to see how these tools would perform on microposts instead of relatively long length texts, for which they have been designed. Some of the evaluated tools have been adjusted to return a restricted set of four named entity types: PER, LOC, ORG and MISC. This was achieved by mapping their result entities to these four classification types. Below is a short description of the evaluated tools.

*ANNIE Named Entity Recognizer* stands for "A Nearly-New IE system" and is a part of the GATE family. ANNIE relies on finite state algorithms, gazetteers and the JAPE language [5]. ANNIE recognizes persons, locations, organizations, dates, addresses and other named entity types.

*Apache OpenNLP*[1] library is a NLP toolkit based on machine learning and maximum entropy models. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution.

*Illinois Named Entity Tagger* is a tagger that tags plain text with named entities with either four label types (person, organization, location and miscellaneous) or 18 label types based on the OntoNotes corpus. It uses gazetteers extracted from Wikipedia, word class models derived from unlabeled text, and expressive non-local features [6].

*Illinois Wikifier* identifies "important expressions" in the input text and cross-links them to Wikipedia. The software is based on a Wikipedia link structure. The structure provides important information about which disambiguations are compatible [7].

*LingPipe*[2] is an NLP tool kit for processing text and performing tasks such as finding the names of people, organizations or locations in news; automatically classifying Twitter search results into categories; or suggesting correct spellings of queries. LingPipe's classification, tagging, and entity extraction are all based on n-gram character language models [8]. There are currently three NER models available for LingPipe: 1. English News: MUC-6, 2. English Genes: GeneTag and English Genomics: GENIA. We have evaluated LingPipe with the first model, which is not suitable for microposts, but was the most appropriate from the models.

*Open Calais*[3] uses NLP and machine-learning techniques to examine the text and locate named entities like people, places,

[1] http://opennlp.apache.org

[2] http://alias-i.com/lingpipe

[3] http://www.opencalais.com/about

products, etc. There are currently 39 entities (e.g. Anniversary, City, Company, Continent) and 83 events and facts (e.g. Acquisition, Alliance, AnalystEarningsEstimate) supported by Calais.

*Stanford Named Entity Recognizer (NER)* (also known as CRFClassifier) is a Java implementation of a Named Entity Recognizer and linear chain Conditional Random Field (CRF) sequence models, coupled with feature extractors for Named Entity Recognition [9]. It is also a part of the Stanford CoreNLP, an integrated suite of NLP tools for English. Stanford NER includes three NER models, each with its caseless version. We have used the four class caseless model (PERSON, LOCATION, ORGANIZATION and MISC), which was trained on CoNLL dataset [10]. When considering the evaluation results, it should be stressed that this model was intended to be used on news texts rather than microposts.

*Wikipedia Miner*[4] is a text annotation tool, which is capable of annotating Wikipedia topics in a given text. We have used this software to discover Wikipedia topics in microposts. Discovered topics were then tagged using a DBPedia Ontology to match the four class named entity set: PER, LOC, ORG and MISC.

## III. Evaluation Dataset

Evaluation dataset has been adopted from Making Sense of Microposts 2013 IE Challenge (further as MSM Challenge). Original MSM Challenge dataset[5] consisted of two parts: training part with 2815 manually annotated microposts and test part with 1526 unannotated microposts. To ensure anonymity, authors of the original dataset have replaced all username mentions with "_MENTION_" and all URLs with "_URL_". There was approximately 4% overlap between both dataset parts including micropost duplicates within each part. We have removed duplicate and overlapping microposts from the training dataset and made it our evaluation dataset. This dataset contained 2752 unique manually annotated microposts with classification restricted to four entity types: 1) Person (PER) - full or partial person names; 2) Location (LOC) - full or partial (geographical or physical) location names, including: cities, provinces or states, countries, continents and (physical) facilities; 3) Organization (ORG) - full or partial organization names, including academic, state, governmental, military and business or enterprise organizations; 4) Miscellaneous (MISC) - a concept not covered by any of the categories above, but limited to one of the entity types: film/movie, entertainment award event, political event, programming language, sporting event and TV show. Occurrence of a particular Named Entity type in the evaluation dataset is depicted in Fig. 1.

## IV. Evaluation Framework

We have chosen GATE [11] as an evaluation framework, because we are familiar with it and because it provides a variety of tools for automatic evaluation. GATE contains an Annotation Diff tool, which compares two annotation sets within a document. Another tool in GATE is Corpus QA (Quality Assurance), which provides functionality for comparing annotation sets over an entire corpus. This tool was the key for obtaining evaluation results.

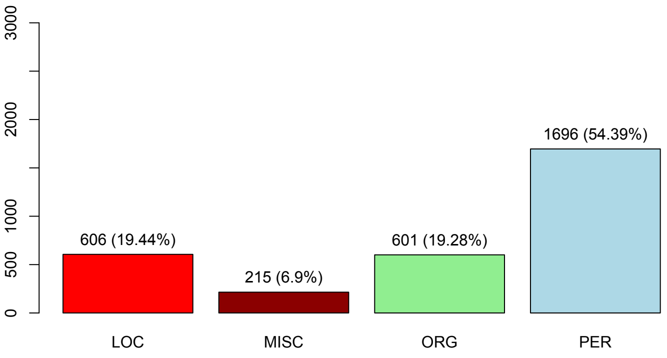

Fig. 1. Named Entity occurrence in the evaluation dataset

Moreover, GATE has a great plug-in system offering an API to develop custom plug-ins, so one can easily integrate third party tools. By following this approach, we have integrated Illinois NER, Illinois Wikifier, Wikipedia Miner and Stanford CoreNLP/NER into GATE. In the same way we have integrated OpenCalais, LingPipe, and OpenNLP tools into GATE.

The evaluation dataset has been imported into GATE as a GATE SerialDataStore corpus, where each micropost was represented as a separate GATE document. There has been a "Manual" annotation set created for each document, where we put all its manual annotations. These manual annotation sets formed a gold standard for the evaluation. The evaluation GATE SerialDataStore corpus is available for download[6].

## V. Evaluation Methodology

The evaluation was started by running all the tools over the GATE SerialDataStore corpus. Each tool produced a new annotation set (named by the tool) for each document, where it put all annotations it has found. Then an Annotation Set Transfer PR (Processing Resource) has been applied to map relevant annotation names to the four target classification types. For instance, Person annotations were renamed to PER, Location to LOC, City to LOC, etc. After that, three main evaluation measures were computed (Precision - $P$, Recall - $R$, and $F_1$) for each evaluated tool and target entity type (PER, LOC, ORG and MISC). These measures were computed using the GATE Corpus QA tool, where the key annotation set was the "Manual" annotation set and the response annotation sets were the ones created by the evaluated tools. Matching of annotations has been done in two modes: strict and lenient. The strict one considered all the partially correct responses as incorrect, while the lenient considered them as correct (e.g. PER - "Mr. Smith" and PER - "Smith" are incorrect by strict matching, but correct by lenient matching). Therefore,

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

[5] http://oak.dcs.shef.ac.uk/msm2013/ie_challenge/MSM2013-CEChallengeFinal.zip

[6] http://ikt.ui.sav.sk/microposts

we show the precision, recall and $F_1$ measures separately for the strict mode ($P_S$, $R_S$ and $F_{1S}$) and for the lenient mode ($P_L$, $R_L$ and $F_{1L}$). We have also computed the average values of these measures combining the strict and the lenient results ($P_A$, $R_A$ and $F_{1A}$). Additionally, we have also provided Macro and Micro summaries: while the Macro summary averages $P$, $R$ and $F_1$ measures on a per document basis, the Micro summary considers the whole dataset as one document.

## VI. Evaluation Results

Tables I to VIII summarize the evaluation results for each tool and each target entity type. Some of the evaluated tools (namely ANNIE, Apache OpenNLP and LingPipe) could not recognize the MISC entity type. For these tools, we did not measure the MISC recognition performance.

Evaluation results of ANNIE can be seen in TABLE I. ANNIE did not hit the highest score in any measure, but it was not the worst in any either. ANNIE performed best within its own measures in LOC classification ($F_{1S}$ = 68%), then in PER ($F_{1S}$ = 61%) and ORG ($F_{1S}$ = 36%) classification. The MISC entity type was not evaluated, because ANNIE does not recognize it.

TABLE I
Evaluation Results of ANNIE

| NE | $P_S$ | $R_S$ | $F_{1S}$ | $P_L$ | $R_L$ | $F_{1L}$ | $P_A$ | $R_A$ | $F_{1A}$ |
|---|---|---|---|---|---|---|---|---|---|
| LOC | 0.77 | 0.60 | 0.68 | 0.80 | 0.63 | 0.71 | 0.79 | 0.62 | 0.69 |
| MISC | - | - | - | - | - | - | - | - | - |
| ORG | 0.42 | 0.31 | 0.36 | 0.48 | 0.36 | 0.41 | 0.45 | 0.33 | 0.38 |
| PER | 0.66 | 0.56 | 0.61 | 0.80 | 0.69 | 0.74 | 0.73 | 0.62 | 0.67 |
| Macro | 0.71 | 0.37 | 0.41 | 0.77 | 0.42 | 0.47 | 0.74 | 0.39 | 0.44 |
| Micro | 0.64 | 0.48 | 0.55 | 0.74 | 0.56 | 0.64 | 0.69 | 0.52 | 0.60 |

TABLE II shows the results of Apache OpenNLP. This tool was not the best in any measure. It was the worst in LOC and ORG recall (strict, lenient and average). But it hit 87% in lenient PER precision, which was the second highest score for this measure among all the evaluated tools. The MISC entity type was not evaluated, since Apache OpenNLP does not classify the entity types subsumed under MISC.

TABLE II
Evaluation Results of Apache OpenNLP

| NE | $P_S$ | $R_S$ | $F_{1S}$ | $P_L$ | $R_L$ | $F_{1L}$ | $P_A$ | $R_A$ | $F_{1A}$ |
|---|---|---|---|---|---|---|---|---|---|
| LOC | 0.66 | 0.41 | 0.51 | 0.72 | 0.45 | 0.56 | 0.69 | 0.43 | 0.53 |
| MISC | - | - | - | - | - | - | - | - | - |
| ORG | 0.31 | 0.24 | 0.27 | 0.39 | 0.30 | 0.34 | 0.35 | 0.27 | 0.30 |
| PER | 0.74 | 0.48 | 0.58 | 0.87 | 0.56 | 0.68 | 0.81 | 0.52 | 0.63 |
| Macro | 0.68 | 0.28 | 0.34 | 0.75 | 0.33 | 0.39 | 0.71 | 0.31 | 0.37 |
| Micro | 0.62 | 0.38 | 0.48 | 0.73 | 0.45 | 0.55 | 0.68 | 0.42 | 0.52 |

Evaluation results of Illinois NER classifier are displayed in TABLE III. Illinois NER performed as the best tool in PER classification ($F_{1S} = 79\%$) and the second best in LOC classification ($F_{1S} = 73\%$). It was also the best in PER recall ($R_S = 78\%$). Illinois NER performed better within its own measures in PER classification ($F_{1S} = 79\%$) than in LOC ($F_{1S} = 73\%$) and ORG ($F_{1S} = 36\%$) classification. Classification of MISC entity type was its weakest one ($F_{1S} = 10\%$).

TABLE III
Evaluation Results of Illinois NER

| NE | $P_S$ | $R_S$ | $F_{1S}$ | $P_L$ | $R_L$ | $F_{1L}$ | $P_A$ | $R_A$ | $F_{1A}$ |
|---|---|---|---|---|---|---|---|---|---|
| LOC | 0.74 | 0.71 | 0.73 | 0.79 | 0.76 | 0.78 | 0.77 | 0.74 | 0.75 |
| MISC | 0.08 | 0.15 | 0.10 | 0.13 | 0.25 | 0.17 | 0.10 | 0.20 | 0.14 |
| ORG | 0.34 | 0.39 | 0.36 | 0.43 | 0.49 | 0.46 | 0.38 | 0.44 | 0.41 |
| PER | 0.79 | 0.78 | 0.79 | 0.84 | 0.83 | 0.84 | 0.82 | 0.80 | 0.81 |
| Macro | 0.49 | 0.51 | 0.50 | 0.55 | 0.58 | 0.56 | 0.52 | 0.55 | 0.53 |
| Micro | 0.60 | 0.65 | 0.62 | 0.66 | 0.71 | 0.68 | 0.63 | 0.68 | 0.65 |

TABLE IV contains the evaluation results of Illinois Wikifier. Illinois Wikifier has achieved the best score in PER precision ($P_S = 84\%$). Within its own measures, this tool performed better in PER classification ($F_{1S} = 63\%$) than in LOC ($F_{1S} = 55\%$), ORG ($F_{1S} = 42\%$) and MISC ($F_{1S} = 16\%$) classification.

TABLE IV
Evaluation Results of Illinois Wikifier

| NE | $P_S$ | $R_S$ | $F_{1S}$ | $P_L$ | $R_L$ | $F_{1L}$ | $P_A$ | $R_A$ | $F_{1A}$ |
|---|---|---|---|---|---|---|---|---|---|
| LOC | 0.57 | 0.53 | 0.55 | 0.62 | 0.58 | 0.60 | 0.59 | 0.56 | 0.57 |
| MISC | 0.13 | 0.22 | 0.16 | 0.15 | 0.26 | 0.19 | 0.14 | 0.24 | 0.17 |
| ORG | 0.41 | 0.42 | 0.42 | 0.46 | 0.47 | 0.47 | 0.44 | 0.45 | 0.44 |
| PER | 0.84 | 0.50 | 0.63 | 0.89 | 0.53 | 0.67 | 0.87 | 0.52 | 0.65 |
| Macro | 0.49 | 0.42 | 0.44 | 0.53 | 0.46 | 0.48 | 0.51 | 0.44 | 0.46 |
| Micro | 0.57 | 0.47 | 0.52 | 0.62 | 0.51 | 0.56 | 0.60 | 0.49 | 0.54 |

LingPipe, which used NER model for news texts, hit the best score in lenient ORG recall (54%), but its precision was the worst among all the evaluated tools (TABLE V). This tool was also the weakest in LOC classification ($F_{1S} = 30\%$). Its low performance was caused by the fact that it used a model for English news texts trained on MUC-6 corpora. It is obvious that it would need to re-train the model on a corpus of labeled microposts.

TABLE V
Evaluation Results of LingPipe

| NE | $P_S$ | $R_S$ | $F_{1S}$ | $P_L$ | $R_L$ | $F_{1L}$ | $P_A$ | $R_A$ | $F_{1A}$ |
|---|---|---|---|---|---|---|---|---|---|
| LOC | 0.26 | 0.52 | 0.35 | 0.30 | 0.59 | 0.40 | 0.28 | 0.56 | 0.37 |
| MISC | - | - | - | - | - | - | - | - | - |
| ORG | 0.04 | 0.25 | 0.07 | 0.09 | 0.54 | 0.15 | 0.06 | 0.40 | 0.11 |
| PER | 0.30 | 0.42 | 0.35 | 0.46 | 0.66 | 0.54 | 0.38 | 0.54 | 0.44 |
| Macro | 0.40 | 0.30 | 0.19 | 0.46 | 0.45 | 0.27 | 0.43 | 0.38 | 0.23 |
| Micro | 0.16 | 0.38 | 0.23 | 0.24 | 0.58 | 0.34 | 0.20 | 0.48 | 0.28 |

Open Calais was the best tool in LOC ($F_{1S} = 74\%$), MISC ($F_{1S} = 27\%$) and ORG ($F_{1S} = 56\%$) classification. It had also the best precision score for these three entity types. Open Calais has placed third in PER classification ($F_{1S} = 69\%$) right after the Illinois NER and Stanford NER. Evaluation results of Open Calais are displayed in TABLE VI.

TABLE VI
EVALUATION RESULTS OF OPENCALAIS

| NE | $P_S$ | $R_S$ | $F_{1S}$ | $P_L$ | $R_L$ | $F_{1L}$ | $P_A$ | $R_A$ | $F_{1A}$ |
|---|---|---|---|---|---|---|---|---|---|
| LOC | 0.80 | 0.68 | 0.74 | 0.83 | 0.71 | 0.77 | 0.82 | 0.70 | 0.75 |
| MISC | 0.38 | 0.21 | 0.27 | 0.57 | 0.31 | 0.40 | 0.47 | 0.26 | 0.34 |
| ORG | 0.73 | 0.45 | 0.56 | 0.79 | 0.49 | 0.60 | 0.76 | 0.47 | 0.58 |
| PER | 0.72 | 0.67 | 0.69 | 0.77 | 0.72 | 0.75 | 0.74 | 0.70 | 0.72 |
| Macro | 0.66 | 0.50 | 0.56 | 0.74 | 0.56 | 0.63 | 0.70 | 0.53 | 0.60 |
| Micro | 0.72 | 0.60 | 0.65 | 0.78 | 0.65 | 0.71 | 0.75 | 0.62 | 0.68 |

Evaluation results of Stanford NER are displayed in TABLE VII. This classifier performed similarly to Illinois NER, which was expected because both tools are based on a sequential prediction and their models were trained on CONLL datasets. Stanford NER was the second best in PER ($F_{1S} = 75\%$) classification after the first Illinois NER ($F_{1S} = 79\%$).

TABLE VII
EVALUATION RESULTS OF STANFORD NER

| NE | $P_S$ | $R_S$ | $F_{1S}$ | $P_L$ | $R_L$ | $F_{1L}$ | $P_A$ | $R_A$ | $F_{1A}$ |
|---|---|---|---|---|---|---|---|---|---|
| LOC | 0.69 | 0.65 | 0.67 | 0.74 | 0.70 | 0.72 | 0.71 | 0.67 | 0.69 |
| MISC | 0.05 | 0.06 | 0.05 | 0.09 | 0.10 | 0.10 | 0.07 | 0.08 | 0.08 |
| ORG | 0.27 | 0.29 | 0.28 | 0.36 | 0.38 | 0.37 | 0.32 | 0.34 | 0.32 |
| PER | 0.75 | 0.74 | 0.75 | 0.83 | 0.82 | 0.82 | 0.79 | 0.78 | 0.78 |
| Macro | 0.44 | 0.44 | 0.44 | 0.50 | 0.50 | 0.50 | 0.47 | 0.47 | 0.47 |
| Micro | 0.59 | 0.59 | 0.59 | 0.66 | 0.66 | 0.66 | 0.62 | 0.63 | 0.63 |

TABLE VIII shows the evaluation results of Wikipedia Miner. We expected a higher score in MISC classification, but this tool performed as the second worst in it ($F_{1S} = 6\%$). At least, we presumed a much higher MISC recall, because we have mapped all the returned annotations which were not in LOC/PER/ORG category to a MISC type. On the other hand, the MISC recall score was the best one ($R_S = 31\%$) as well as LOC ($R_S = 73\%$) and ORG ($R_S = 47\%$).

TABLE VIII
EVALUATION RESULTS OF WIKIPEDIA MINER

| NE | $P_S$ | $R_S$ | $F_{1S}$ | $P_L$ | $R_L$ | $F_{1L}$ | $P_A$ | $R_A$ | $F_{1A}$ |
|---|---|---|---|---|---|---|---|---|---|
| LOC | 0.34 | 0.73 | 0.46 | 0.36 | 0.78 | 0.49 | 0.35 | 0.75 | 0.48 |
| MISC | 0.03 | 0.31 | 0.06 | 0.05 | 0.40 | 0.08 | 0.04 | 0.36 | 0.07 |
| ORG | 0.21 | 0.47 | 0.29 | 0.22 | 0.50 | 0.30 | 0.21 | 0.48 | 0.29 |
| PER | 0.64 | 0.59 | 0.61 | 0.69 | 0.63 | 0.66 | 0.67 | 0.61 | 0.64 |
| Macro | 0.31 | 0.52 | 0.35 | 0.33 | 0.58 | 0.38 | 0.32 | 0.55 | 0.37 |
| Micro | 0.29 | 0.57 | 0.39 | 0.31 | 0.62 | 0.42 | 0.30 | 0.59 | 0.40 |

For easier comparison of the tools' performance over the dataset, we present several plots below. In each, we consider only strict-match-computed measures, i.e. $P_S$, $R_S$ and $F_{1S}$. Precision measures of all the evaluated tools over the dataset are depicted in Fig. 2.

The highest precision in LOC, MISC and ORG entity type was achieved by Open Calais (80%, 38% and 73% respectively). The highest precision in PER entity type was measured for Illinois Wikifier (84%) followed by Illinois NER (79%) and Stanford NER (75%).

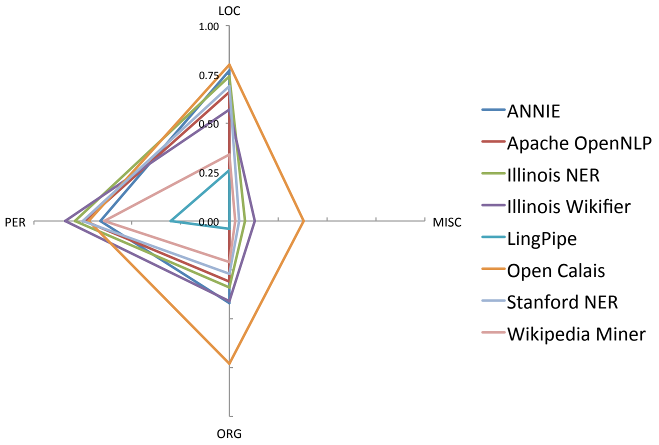

Fig. 2. Precision of evaluated tools by entity type

Recall measures of all the evaluated tools over the dataset are depicted in Fig. 3. The highest recall in LOC, MISC and ORG entity type was achieved by Wikipedia Miner (73%, 31% and 47% respectively). The highest recall in PER entity type was measured for Illinois NER classifier (78%) followed by Stanford NER (74%) and Open Calais (67%).

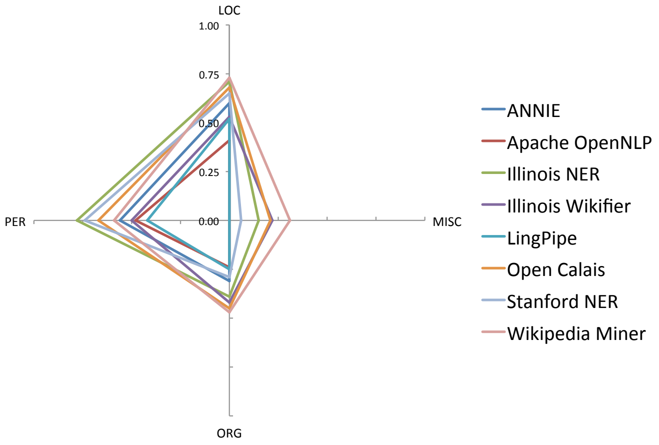

Fig. 3. Recall of evaluated tools by entity type

$F_1$ measures of all the evaluated tools over the dataset are depicted in Fig. 4. The best $F_1$ score in LOC, MISC and ORG classification was measured for Open Calais (74%, 27% and 56% respectively), while the best $F_1$ score in PER classification had Illinois NER (79%) followed by Stanford NER (75%) and Open Calais (69%). Illinois NER was the second best in LOC classification (73%) followed by the third ANNIE (68%). The second and the third best in MISC classification were Illinois Wikifier (16%) and Illinois NER (10%), respectively. Illinois Wikifier was the second in ORG classification (42%) followed by Illinois NER and ANNIE sharing the third place (36%).

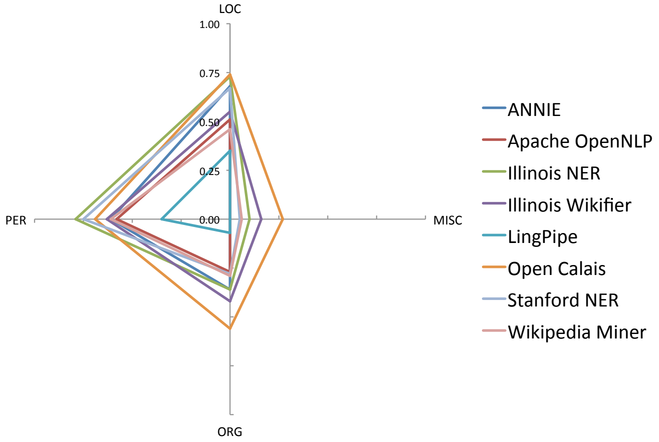

Fig. 4. $F_1$ of evaluated tools by entity type

Dispersion of the $F_1$ score by entity type among the tools is displayed in Fig. 5. Micro summary of the evaluated tools can be seen in Fig. 6 where all the three kinds of $P$, $R$ and $F_1$ measures are plotted out.

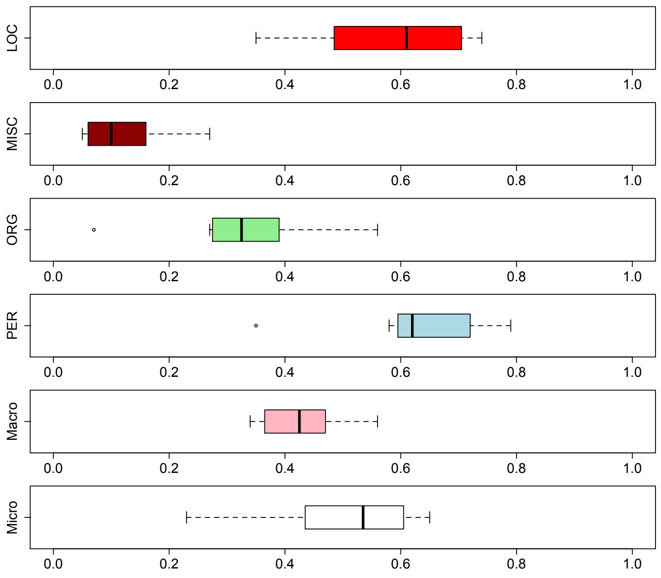

Fig. 5. $F_1$ spread of evaluated tools over the dataset

## VII. Current and Future work

The evaluation results have shown that classifiers return diverse results, but when properly combined, this might lead to a new composite classifier that performs better than any individual classifier on its own. For example, a simple composite classifier might incorporate the best performing tool for each named entity class. The diversity of the results can be seen in Fig. 7. If we make a union of all the entities recognized by the evaluated tools, we get a very high recall, but with a drawback of a very low precision. Machine learning techniques such as decision trees can help us to choose appropriate technique in a specific context and thus improve the precision, while keeping the recall relatively high. Currently, we are working on a NER classifier, which aggregates some of the evaluated tools. To be more concrete, we accumulate various text features as well as NER and POS-tagging results and use them for producing a decision tree model for the classifier. This approach is giving us a higher $F_1$ score than a simple composite classifier. Our observations have also shown that, on average, every second micropost contains a link, which can be an external source of additional information. Such information could be exploited in the NER process, for example in entity disambiguation.

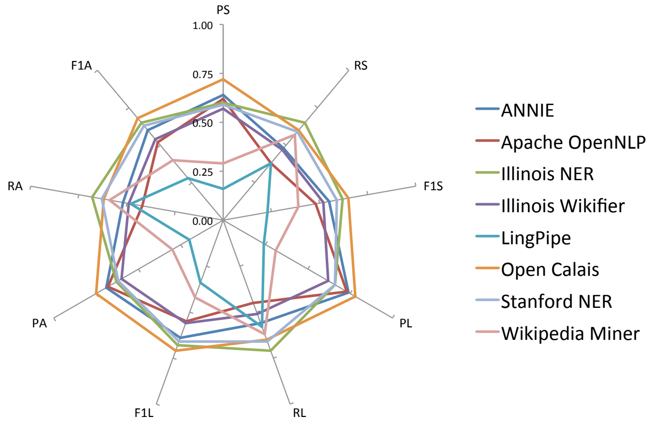

Fig. 6. Micro summary of evaluated tools over the dataset

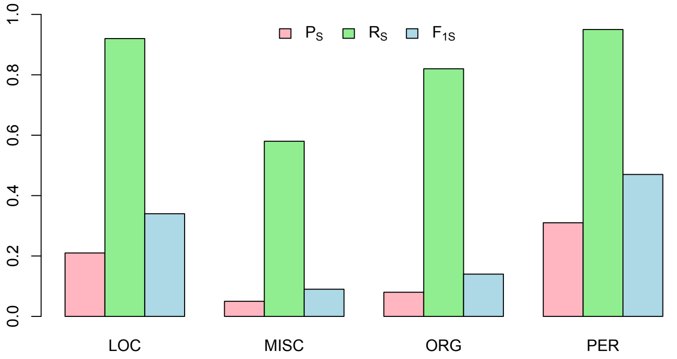

Fig. 7. Performance of unified evaluated tools

## VIII. Conclusion

In this paper we survey and evaluate the state of the art information extraction tools on the same dataset of microposts. These tools were executed without prior tuning for this kind of text and there was no preprocessing applied on microposts. The evaluation gave us valuable information on how these tools behave on microposts and how they perform regarding different NE types. Our findings revealed that the best performing NER tool on microposts was Open Calais; however, our experiment also showed that the total recall of all the tools combined together was much higher. This means that each tool is good at discovering different kinds of entities and that there is a place for combining these tools in order to achieve superior performance.

## Acknowledgment

This work is supported by projects TRA-DICE APVV-0208-10, VEGA 2/0185/13, and VENIS FP7-284984.

## References

[1] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: an experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1524–1534. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145595

[2] B. Locke and J. Martin, "Named entity recognition: Adapting to microblogging," *University of Colorado*, 2009.

[3] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, January 2007, publisher: John Benjamins Publishing Company.

[4] L. Reeve and H. Han, "Survey of semantic annotation platforms," in *Proceedings of the 2005 ACM symposium on Applied computing*, ser. SAC '05. New York, NY, USA: ACM, 2005, pp. 1634–1638. [Online]. Available: http://doi.acm.org/10.1145/1066677.1067049

[5] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[6] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, ser. CoNLL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 147–155. [Online]. Available: http://dl.acm.org/citation.cfm?id=1596374.1596399

[7] L. Ratinov, D. Roth, D. Downey, and M. Anderson, "Local and global algorithms for disambiguation to wikipedia," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1375–1384. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002642

[8] B. Carpenter and B. Baldwin, "Text analysis with lingpipe 4," 2011. [Online]. Available: http://alias-i.com/lingpipe-book/lingpipe-book-0.5.pdf

[9] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[10] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 363–370. [Online]. Available: http://dx.doi.org/10.3115/1219840.1219885

[11] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, *Text Processing with GATE (Version 6)*, 2011. [Online]. Available: http://tinyurl.com/gatebook