Information Extraction for Social Media

Mena B. Habib
Chair Databases
University of Twente
m.b.habib@ewi.utwente.nl

Maurice van Keulen
Chair Databases
University of Twente
m.vankeulen@utwente.nl

Abstract

The rapid growth in IT in the last two decades has led to a growth in the amount of information available online. A new style for sharing information is social media. Social media is a continuously instantly updated source of information. In this position paper, we propose a framework for Information Extraction (IE) from unstructured user generated contents on social media. The framework proposes solutions to overcome the IE challenges in this domain such as the short context, the noisy sparse contents and the uncertain contents. To overcome the challenges facing IE from social media, State-Of-The-Art approaches need to be adapted to suit the nature of social media posts. The key components and aspects of our proposed framework are noisy text filtering, named entity extraction, named entity disambiguation, feedback loops, and uncertainty handling.

1 Introduction

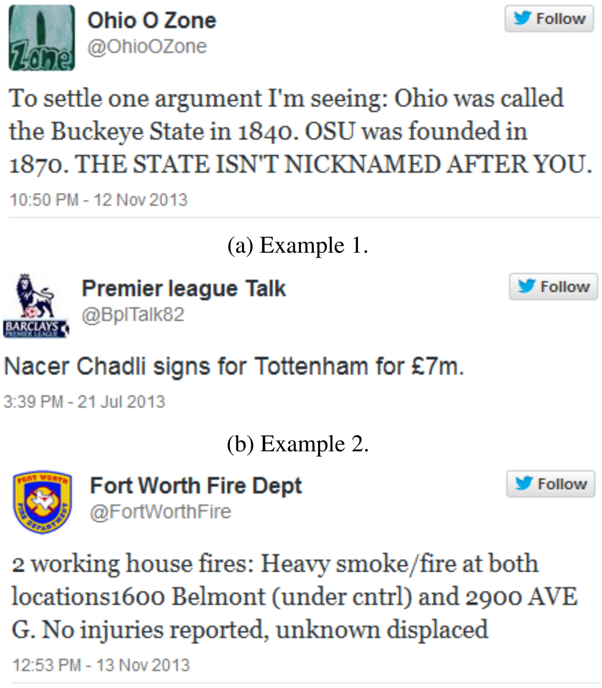
The rapid growth in IT in the last two decades has led to a growth in the amount of information available on the World Wide Web. A new style for exchanging and sharing information is social media. Social media refers to the means of interaction among people in which they create, share, and exchange information and ideas in virtual communities and networks (like Twitter and Facebook). According to CNN[1], more Americans get their news from the Internet than from newspapers or radio, and three-fourths say they hear of news via e-mail or updates on social media sites. Social media, in many cases, provide more up-to-date information than conventional sources like online news. To make use of this vast amount of information, it is required to extract structured information out of these heterogeneous unstructured information. Information Extraction (IE) is the research field that enables the use of such a vast amount of unstructured distributed information in a structured way. IE systems analyse human language text in order to extract information about different types of events, entities, or relationships. Structured information could be stored in Knowledge-bases (KB) which hold facts and relations extracted from the free style text. A KB is an information repository that provides a means for information to be collected, organized, shared, searched and utilized. It can be either machine-readable or intended for human use.

In this paper, we introduce a framework for IE from unstructured user generated contents on social media. Although IE is a field of research that has been studied for long time, there is very few work done on that field for social media contents. (Bontcheva et al., 2013) proposed TwitIE, an open-source NLP pipeline customised to microblog text. However, TwitIE doesn't provide mechanisms for messages filtering or named entity disambiguation or relation/fact extraction. All efforts on IE field focus on facts extraction from encyclopaedias like Wikipedia (Suchanek et al., 2007; Auer and Lehmann, 2007), or from web pages (Nakashole et al., 2011; Carlson et al., 2010; Kushmerick et al., 1997; Crescenzi et al., 2001).

IE from text is an important task in text mining. The general goal of information extraction is to discover structured information from unstructured or semi-structured text. For example, given the tweets shown in figure 1, we can extract the following information:

[1] http://edition.cnn.com/2010/TECH/03/01/social.network.news/index.html

Proceedings of Third Workshop on Semantic Web and Information Extraction, pages 9–16,
Dublin, Ireland, 24 August, 2014.



Ohio O Zone
@OhioOZone

To settle one argument I'm seeing: Ohio was called the Buckeye State in 1840. OSU was founded in 1870. THE STATE ISN'T NICKNAMED AFTER YOU.

10:50 PM - 12 Nov 2013

(a) Example 1.

Premier league Talk
@BplTalk82

Nacer Chadli signs for Tottenham for £7m.

3:39 PM - 21 Jul 2013

(b) Example 2.

Fort Worth Fire Dept
@FortWorthFire

2 working house fires: Heavy smoke/fire at both locations1600 Belmont (under cntrl) and 2900 AVE G. No injuries reported, unknown displaced

12:53 PM - 13 Nov 2013

(c) Example 3.

Figure 1: Tweets examples

Example (1):
Called(U.S. state of Ohio, Buckeye State),
FoundedIn(The Ohio State University, 1870).

Example (2):
SignedFor(Nacer Chadli (the football player), Tottenham Hotspur Football Club).

Example (3):
Fire(1600 Belmont Avenue, Fort Worth, TX),
Fire(2900 Avenue G., Fort Worth, TX).

As we can see in the examples, IE can be applied for open or closed domain. Open IE is to extract all possible relations and facts stated in a post as in examples 1 and 2. Closed domain IE is to extract facts for a specific target domain or fill in predefined templates like example 3. Other meta data could be extracted like the time or the source of the extracted fact. This could help in improving the precision of the extraction process. For instance, in the 3rd example, it is not stated where exactly is the “1600 Belmont Avenue” or “2900 Avenue G.”. We could infer this extra knowledge from the source of the tweet “Fort Worth Fire Dept”. Same with example 2, the word “Tottenham” is ambiguous. Further information about the entity “Nacer Chadli” should help to link “Tottenham” to “Tottenham Hotspur Football Club”.

2 Challenges

Application of the State-Of-The-Art approaches on social media is not reasonable for the following challenges:

- **Informal language:** Posted texts are noisy and written in an informal setting, include misspellings, lack punctuation and capitalisation, use non-standard abbreviations, and do not contain grammatically correct sentences. Traditional KB construction approaches rely mostly on capitalization and

Part-Of-Speech tags to extract the named entities. The lack of such features in social media posts makes the IE task more challenging.

- **Short context:** There is a post length limit on some social media networks like Twitter. This limit forces the users to use more abbreviations to express more information in their posts. The shortness of the posts makes it more challenging to disambiguate mentioned entities and to resolve co-references among tweets.

- **Noisy sparse contents:** The users' posts on social media are not always important nor contain useful information. Around 40% of twitter messages content are pointless babble[2]. Filtering is a pre-processing step that is required to purify the input posts stream.

- **Information about non-famous entities:** The IE State-Of-The-Art approaches link the entities involved in the extracted information to a KB. However, people normally use social media to express information about themselves or about some small local events (street festival or accident) and thus the involved entities are not contained in a KB. New ways of entity linkage need to be introduced to suit IE from social media posts.

- **Uncertain contents:** Of course not every available information is trustworthy. In addition to errors that may take place during the IE process, information contained in users' contributions is often partial, subject to evolution over time, in conflict with other sources, and sometimes untrustworthy. It is required to handle the uncertainty involved in the extracted facts.

3 The State-Of-The-Art

In order to extract information from text, a set of subtasks has to be applied on the input text. Figure 2 shows the subtasks modules of a traditional IE system. Those modules are described according to the State-Of-The-Art IE approaches as follows:

- **Named Entity Extraction:** A named entity is a sequence of words that designates some real world entity (e.g. "California", "Steve Jobs" and "Apple Inc."). The task of named entity extraction (NEE), is to identify named entities from free-form text. This task cannot be simply accomplished by string matching against pre-compiled gazetteers because named entities of a given entity type usually do not form a closed set and therefore any gazetteer would be incomplete. NEE approaches mainly use capitalization features and Part-Of-Speech tags for recognizing named entities. Part-Of-Speech (POS) tagging is the process of marking up a word in a text (corpus) as corresponding to a particular Part-Of-Speech, based on both its definition, as well as its context (i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph). A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

- **Named Entity Disambiguation:** In natural language processing, named entity disambiguation (NED) or entity linking is the task of determining the identity of entities mentioned in text. For example, to link the mention "California" to the Wikipedia article "`http://en.wikipedia.org/wiki/California`". It is distinct from named entity extraction (NEE) in that it identifies not the occurrence of names but their reference. NED needs a KB of entities to which names can be linked. A popular choice for entity linking on open domain text is Wikipedia (Cucerzan, 2007; Hoffart et al., 2011).

- **Fact Extraction:** In open IE, the goal of the fact extraction (FE) module is to detect and characterize the semantic relations between entities in text or relations between entities and values. In closed domain IE, the goal is to fill in a predefined template using the extracted named entities.

[2] `http://web.archive.org/web/20110715062407/www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf`

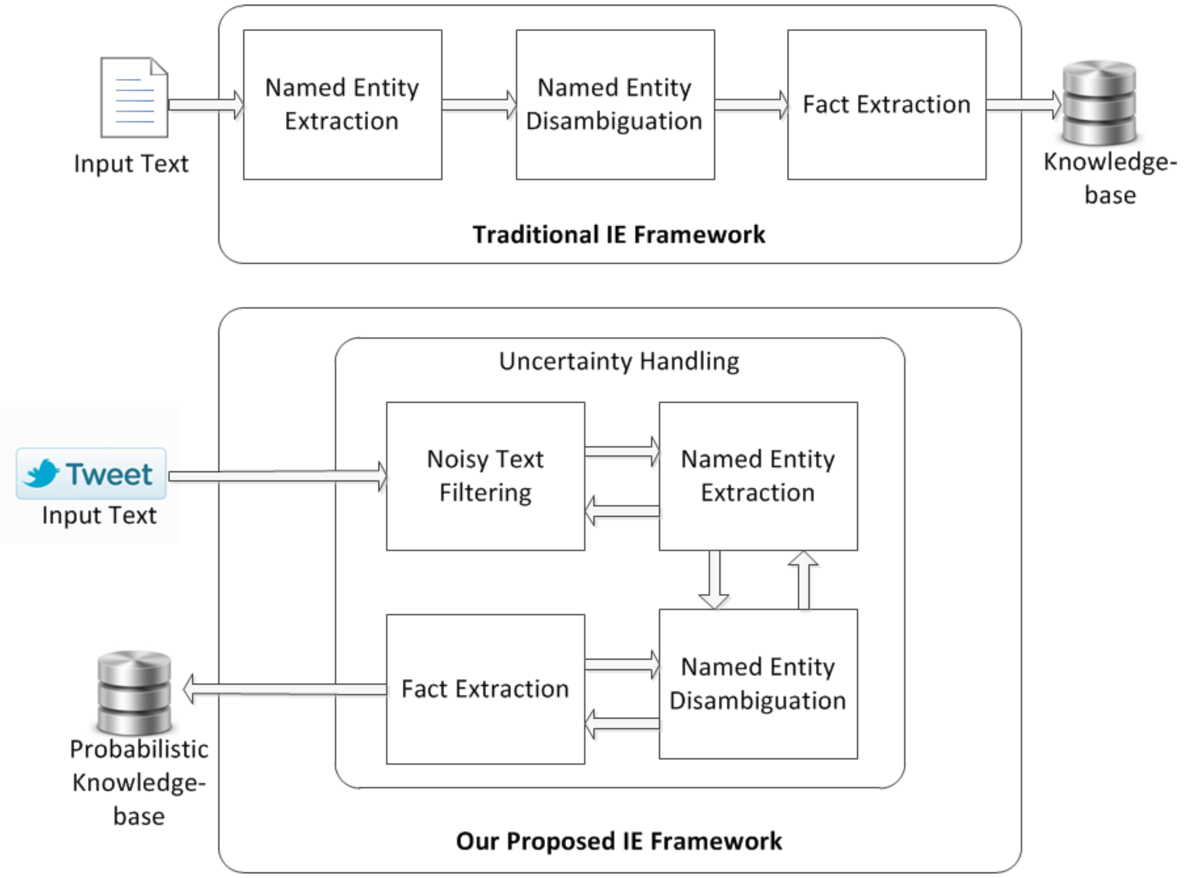


Figure 2: Traditional IE framework versus our proposed IE framework.

4 Proposed Framework

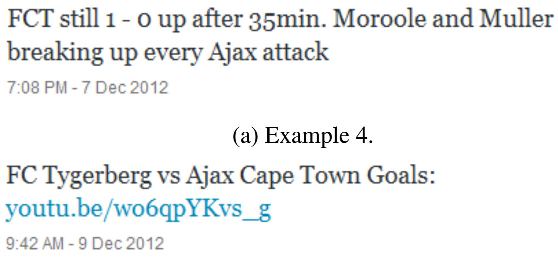
To overcome the challenges facing IE from social media, State-Of-The-Art approaches need to be adapted to suit the nature of social media posts. Here, we describe the key components and aspects of our proposed framework (see figure 2) and show how it would overcome the challenges.

- **Noisy Text Filtering:** There are millions of social media posts every day. For example, the average number of tweets exceeds 140 million tweet per day sent by over 200 million users around the world. These numbers are growing exponentially[3]. This huge number of posts not always contains useful information about users, locations, events, etc. It is required to filter non-informative posts. Filtering could be done based on domain or language or other criteria to make sure to keep only relevant posts that contains information about the domain need to be processed. For example, if we want to extract the results of all the football World Cup matches from tweets, we need to filter millions of tweets to get only the subset of tweets that contain information about results of matches, note that even this subset may contains predicted results or results changing during the matches.

- **Named Entity Extraction:** With the lack of formal writing style, we need new approaches for NEE that don't rely heavily on syntactic features like capitalization and POS. In (Habib et al., 2013), we participated in a challenge to extract named entities from microposts of tweets, we proposed a new approach that combines State-Of-The-Art techniques with clues derived from disambiguation step to detect named entities. Our system named to be the best among all the challenge participants (Basave et al., 2013).

- **Named Entity Disambiguation:** As stated in the State-Of-The-Art section, researchers normally link entities to Wikipedia articles or to KB entries. For social media posts, sometimes this is not

[3] http://www.marketinggum.com/twitter-statistics-2011-updated-stats/



FCT still 1 - 0 up after 35min. Moroole and Muller breaking up every Ajax attack

7:08 PM - 7 Dec 2012

(a) Example 4.

FC Tygerberg vs Ajax Cape Town Goals: youtu.be/wo6qpYKvs_g

9:42 AM - 9 Dec 2012

(b) Example 5.

Figure 3: Tweets examples

possible as many of the mentioned entities cannot be linked to Wikipedia articles or a KB entries. However, normally users have home pages or profiles on a social media network. Furthermore, festivals and local events also commonly have home pages representing these events. In (Habib and van Keulen, 2013), we proposed an open world approach for NED for tweets. Named entities are disambiguated by linking them to a home page or a social network profile page in case they don't have a Wikipedia article. Target tweets (tweets revolving around same event) are used to enrich the tweet context and hence to improve the effectiveness of finding the correct entity page. Other meta data from users profiles could also be used to improve the disambiguation process.

- **Feedback Loops:** In figure 2, we can see, in the traditional IE framework, the pipeline of the subtasks. Each subtask processes the input and generates an output and passes this output to the next subtask. There is no possibility of modifying or refining the output of one subtask once it is already generated. In our framework, feedback plays a key role in the system. Every subtask gives a feedback to the preceding subtask which allows for possibility of iterations of refinement (Habib and van Keulen, 2012). For example, if the NEE module extracted the mention "Apple". And when NED module tries to disambiguate the extracted mention, it finds that it could not be linked to any entity. This means that most probably this mention "Apple" refers to the fruit rather than the company. In traditional approaches, such feedback cannot be passed, and the NED has to find a page to link the extracted mention anyway. Furthermore, as "Apple" is not considered a named entity anymore this may affect the decision made that this piece of text in non-informative and thus should be filtered. This is typically how human beings interpret text. In (Habib et al., 2014), we applied the proposed feedback loop on the #Microposts 2014 Named Entity Extraction and Linking Challenge. Our system is ranked second among all the challenge participants (Cano Basave et al., 2014).

  Similarly, the feedback loop takes place between the FE and the NED modules. This feedback helps resolving errors that took place earlier in the disambiguation step. For example in figure 3a, one might interpret that the tweet refers to a match of "FC Twente" versus "Ajax Amsterdam" in the Dutch football league. Unfortunately, this turns to be a wrong assumption after checking the tweet in figure 3b which shows that the match was between "FC Tygerberg" and "Ajax Cape Town" in the South African second division football league. A feedback from the FE module should trigger and correct the wrong decision made earlier in the NED module. It is also possible that the FE module sends a feedback message to the noisy text filtering module that the message is non-informative if it failed to extract the required information or if the extracted fact contradicts other facts or rules. For example, if we want to extract facts about the football World Cup, and we found a tweet the contains a fact about football club (not national team) then a feedback message is sent back to the noisy text filtering module to mark this tweet as irrelevant one.

- **Uncertainty Handling:** As mentioned in the challenges, the information contained in the social media posts involves high degree uncertainty due to many reason. We envision an approach that fundamentally treats annotations and extracted information as uncertain throughout the process.

(Goujon, 2009) models this uncertainty in a fuzzy way, however we believe that a probabilistic approach would be a better solution to handle such uncertainty. Probabilistic knowledge-bases (PKB) are KBs where each fact is associated with a probability indicating how trustworthy is this fact. Probabilities are updated according to many factors like time, users, contradiction or compatibility with other facts, etc.

Using the same example (figure 3a) mentioned above, the mention "FCT" is linked to "FC Twente" with some certainty confidence. This probability should be adjusted after processing the second tweet shown in figure 3b which holds a contradicting fact about the mention "FCT". Furthermore, a new fact is added to the KB indicating that "FCT" is linked to "FC Tygerberg". The benefit of using a PKB is that we can keep both interpretations "FC Twente" and "FC Tygerberg" with different probabilities assigned to them. Using a PKB, all information is preserved.

Another source of uncertainty is the knowledge updating. One true fact at certain point of time may be wrong at a later point of time. Scores of sport games change over time. Twitter users normally tweet about the score during and after the game. They may also write their predictions on the game prior to the game itself. A probabilistic model should be developed to handle those uncertainties using evidences like number of tweets with the same extracted result, number of re-tweets, time of the tweets, last extracted result about the game, etc.

- **Modules Portability:** Each module from our proposed framework could be customized and reused individually or embedded inside other frameworks. For example, NEE and NED modules could be used in a sentiment analysis system that measures the users opinions towards some product. Noisy text filtering could be embedded inside a search engine for social media posts.

5 Knowledge exchange and impact

The aim of this position paper is to propose a framework for information extraction from unstructured user generated contents on social media. IE systems analyse human language text in order to extract information about different types of events, entities, or relationships. Structured information could be stored in KB which hold facts and relations extracted from the free style text. A KB is a special kind of database for knowledge management. A KB is an information repository that provides a means for information to be collected, organized, shared, searched and utilized. Information extraction has applications in a wide range of domains. There is many stakeholders that would benefit from such framework. Here, we give some examples for applications of information extraction:

- Financial experts always look for specific information to help their decision making. Social media is a very important source of information about shareholders attitudes and behaviours. For example, a finance company may need to know the shareholders reaction towards some political action. Automatically finding such information from users posts on social media requires special information extraction technologies to analyse social media streams and capture such information at runtime.

- Security agencies normally analyse large amounts of text manually to search for information about people involved in criminal or terrorism activities. Social media is a continuously instantly updated source of information. Football hooligans sometimes start their fight electronically on social media networks even before the sport event. This information could be helpful to take actions to prevent such violent, and destructive behaviours.

- With the fast growth of the Web, search engines have become an integral part of people's daily lives, and users' search behaviours are much better understood now. Search based on bag-of-word representation of documents provides less satisfactory results for the new challenges and demands. More advanced search problems such as entity search, and question answering can provide users with better search experience. To facilitate these search capabilities, information extraction is often needed as a pre-processing step to enrich document representation or to populate an underlying database.

Our main goal of this proposal is to provide an open source set of portable and customizable modules that can be used by different stakeholders with different application needs on social media contents. Open source software is a computer software with its source code made available and licensed with a license in which the copyright holder provides the rights to study, change and distribute the software to anyone and for any purpose. This enables the ICT community from not only using but also developing and extending the system according to their needs. Individuals and organizations always choose open source software for their zero cost, and its adaptability.

Reusability would be a key feature in our framework design. In software industry, reusability is the likelihood that a part of a system can be used again to add new functionalities with slight or no modification. Reusable modules reduce implementation time and effort. As an example for possible contribution to the society, we contribute to the TEC4SE project [4]. The aim of the project is to improve the operational decision-making within the security domain by gathering as much information available from different sources (like cameras, police officers on field, or social media posts). Then these information is linked and relationships between different information streams are found. The result is a good overview of what is happening in the field of security in the region. Our contribution to this project to filter twitter stream messages and enrich it by extracting named entities at run time. It will be more valuable to this project to complete the whole IE process by building a complete KB from the extracted information for further or later investigations.

6 Conclusion

IE for social media is an emerging field of research. The noisy contents, shortness of posts, informality of used language, and the uncertainty involved, add more challenges to IE for social media over those of formal news articles. In this paper we propose a framework to cope with those challenges through set of portable modules. Messages filtering, feedback loops, and uncertainty handling are the key aspects of our framework.

References

Sören Auer and Jens Lehmann. 2007. What have innsbruck and leipzig in common? extracting semantics from wiki content. In *Proceedings of the 4th European Conference on The Semantic Web: Research and Applications*, ESWC '07, pages 503–517, Berlin, Heidelberg. Springer-Verlag.

Amparo E. Cano Basave, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie, editors. 2013. *Proceedings, Concept Extraction Challenge at the 3rd Workshop on Making Sense of Microposts (#MSM2013): Big things come in small packages, Rio de Janeiro, Brazil, 13 May 2013*, May.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark Greenwood, Diana Maynard, and Niraj Aswani. 2013. Twitie: An open-source information extraction pipeline for microblog text. In *In Proceedings of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics*.

Amparo Elizabeth Cano Basave, Giuseppe Rizzo, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. 2014. Making sense of microposts (#microposts2014) named entity extraction & linking challenge. In *4th Workshop on Making Sense of Microposts (#Microposts2014)*, pages 54–60.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., and Tom M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 101–110, New York, NY, USA. ACM.

Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. 2001. Roadrunner: Towards automatic data extraction from large web sites. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pages 109–118, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June. Association for Computational Linguistics.

[4] `http://www.tec4se.nl/`

Bénédicte Goujon. 2009. Uncertainty detection for information extraction. In *RANLP*, pages 118–122.

Mena B. Habib and Maurice van Keulen. 2012. Improving toponym disambiguation by iteratively enhancing certainty of extraction. In *Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval, KDIR 2012, Barcelona, Spain*, pages 399–410, Spain, October. SciTePress.

Mena B. Habib and Maurice van Keulen. 2013. A generic open world named entity disambiguation approach for tweets. In *Proceedings of the 5th International Conference on Knowledge Discovery and Information Retrieval, KDIR 2013, Vilamoura, Portugal*, pages 267–276, Portugal, September. SciTePress.

Mena B. Habib, Maurice Van Keulen, and Zhemin Zhu. 2013. Concept extraction challenge: University of Twente at #msm2013. In Basave et al. (Basave et al., 2013), pages 17–20.

Mena B. Habib, Maurice van Keule, and Zhemin Zhu. 2014. Named entity extraction and linking challenge: University of twente at #microposts2014. In *4th Workshop on Making Sense of Microposts (#Microposts2014)*, pages 64–65.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nicholas Kushmerick, Daniel S. Weld, and Robert Doorenbos. 1997. Wrapper induction for information extraction. In *Proc. IJCAI-97*.

Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. 2011. Scalable knowledge harvesting with high precision and high recall. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 227–236, New York, NY, USA. ACM.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.