Handling Uncertainty in Relation Extraction: A Case Study on Tennis Tournament Results Extraction from Tweets

Jochem Verburg
University of Twente
j.g.j.verburg@student.utwente.nl

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

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

ABSTRACT

Relation extraction involves different types of uncertainty due to the imperfection of the extraction tools and the inherent ambiguity of unstructured text. In this paper, we discuss several ways of handling uncertainties in relation extraction from social media. Our study case is to extract tennis games' results for two Grand Slam tennis tournaments from tweets. Analysis has been done to find to what extent it is useful to use semantic web, domain knowledge, facts repetition, and authors' trustworthiness to improve the certainty of the extracted relations.

1. INTRODUCTION

Five hundred million tweets are sent every day [13]. To make use of this vast amount of information, it is required to extract structured information out of this heterogeneous unstructured information. However, Information Extraction (IE) from tweets is challenging due to the various sources of uncertainty. 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. In this paper, we investigate different sources of uncertainty and propose methods to improve the certainty of the extracted relations. To validate our methods, we used a case study where the results of two Grand Slam tennis tournaments are extracted from tweets.

1.1 Case description

In this paper, we use two self-collected tweets' datasets about Roland Garros 2014 (RG) and Wimbledon 2014 (WI), with the aim to extract the games' results. The tennis domain is useful as a case study for several reasons. First, due to its popularity, a large number of tweets are posted during the big tournaments. Furthermore, most of the mentioned entities are covered in knowledge-bases (KB) like DBpedia and Yago, giving the opportunity not only to extract and disambiguate entities but also to extract (defeat) relations, like in [3]. In this case, large-scale automatic analysis and validation is possible since the ground truth for the relations are available. Last,

K-Cap '15 Palisades, NY USA

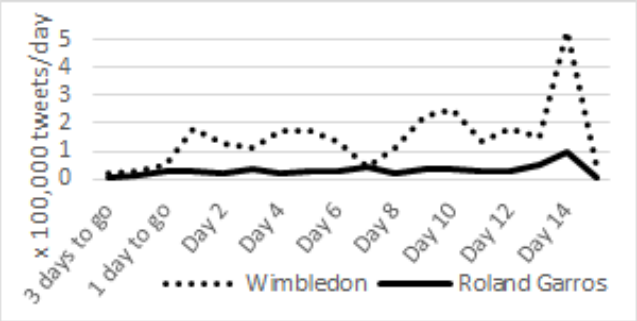

Figure 1: Number of tweets/day for each tournament

it is a domain which has the advantage over other sport domains like football, that most of the matches are tweeted about just a few times. For example, Van Oorschot et al. [14] had problems in their event detection in less popular games.

The datasets have been collected from Twitter using a keyword-search: '*Wimbledon*', '*rolandgarros*' and '*#rg14*'. Tweets in all languages were collected but only the English ones are considered in our research.

One major difficulty in IE from tweets is that the short messages often lack context [7]. However, as we have a closed domain, we can safely assume that if a mention of a tennis player is extracted, it can be linked directly to the entity of the tennis player without spending much effort on the disambiguation process. The usefulness of hashtags has already been tested by Adedoyin-Olowe et al. [1], Wagner and Strohmaier [15], and others, and a keyword-search is in essence not different. The WI-dataset and the RG-dataset contain 2,683,270 and 584,400 English tweets respectively. The distribution can be found in Figure 1.

2. EXTRACTION METHOD

For IE, we use a modular approach. The module used in each step can be replaced by another implementation/approach. The modules are: *a*) Named Entity Recognition (NER); *b*) Named Entity Linking (NEL); and *c*) Relation Extraction (RE).

For the first, the Stanford NER [5] is used to extract named entities of type *Person* from the tweets. Alternatives which focus on NER for tweets are available. However, we preferred Stanford NER due to its efficiency. For example, the system of Ritter et al.[12] takes 2.04 seconds/tweet on a set of 10,000 tweets whereas Stanford NER needed only 0.00163 seconds/tweet. Stanford NER was used to do more analyses on a large scale. The output of the NER module is used as an input for NEL module, as explained later in

Table 1: Initial results Relation Extraction

	Precision	Recall	F_1-measure
Wimbledon	0.426	0.665	0.520
Max. WI	1.000	0.914	0.955
Roland Garros	0.503	0.681	0.579
Max. RG	1.000	0.878	0.935

this section.

The semantic web is used to link (disambiguate) NEs belonging to a specific subclass (for example a male participant of WI, is a subclass of Person). The subclass can be generated by making a set of entities using a list of names. NEL is done by taking all entities classified as the superclass and then checking for a (partial) match with the names of entities in the subclass. These links also make more normalization possible, since all identified entities can be replaced by their URI, which is something the method of Liu et al. [11] still has difficulties with. By using the semantic web, no training data with manually added ground truth is needed to extract NEs for a specific domain. This in contrast to the NER method proposed by Ritter et al. [12] which still needs new training. The method is tested with two data sources in the semantic web: DBPedia[1] and Yago [8]. Yago will be included in section 3.4 to check how using a different knowledge source influences the results.

For the NEL process, a domain specific KB is constructed by matching the players' names taken from the corresponding websites of WI[2] and RG[3] to the entries of tennis players of the KB. Then this domain specific KB is used to match names in the text (if a name of a person is part of one of the names of the entity, we consider it as a match). This method does not look at other entities in the semantic web, since it is assumed that the data set restricts the domain enough. This means, that if there would be an entry in the KB in which 'Djokovic' is a football player, 'Djokovic' in the text of our tweets would still be linked to the tennis player entry.

The last step is the defeat-RE. This is done for both men's and women's singles. Regular expressions are used for the extraction due to its simplicity. First, a regular expression is made to match extracted 'person' entity against the players' names in the tweet (Expression 1). Then, expression 2 is used to find all defeat-relations.

$$personRegex = (entity1|entity2|entity...) \quad (1)$$

$$personRegex((?!personRegex).)^* (beat|defeat|def.|wins).^*?personRegex \quad (2)$$

The evaluation of the RE is done by comparing the extracted relations against the ground truth of the games' results. In the ground truth, all players were also automatically linked to the corresponding entries of tennis players. The initial results using DBPedia can be found in Table 1. The mentioned maximum results are the upper bound results that we can not exceed. They are found by checking for every result in the ground truth, whether both players are mentioned together in at least one tweet. This leads to the maximum possible recall. The maximum precision is theoretically 1.

[1] http://dbpedia.org

[2] http://www.wimbledon.com/

[3] http://www.atpworldtour.com/Share/Event-Draws.aspx?Year=2014\&EventId=520\&Draw=ms

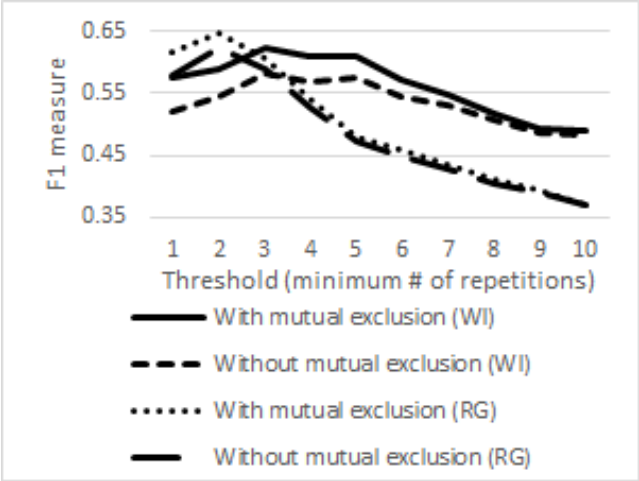

Figure 2: F_1 measures mutual exclusion

3. HANDLING UNCERTAINTY

Uncertainties in IE come from both the data sources and the extraction methods [9] and can be caused at every level. For example, Habib and Van Keulen [6] mentioned that the errors in the output of the NER module propagates to the NEL module. If tennis players are not recognized as a Person by the NER module, then it won't be analyzed further on. If the NEL links a name to the wrong player, then the RE will contain mistakes. Lastly, the RE method itself can also extract wrong relations.

Extensive research has already been done on NER for tweets [12]. Techniques for RE have also been researched [4] and [3]. Therefore, here we focus on the analysis of the uncertainty caused by data sources. In future research, we want to investigate the uncertainties resulting from extraction methods.

3.1 Mutual exclusiveness

Inspired by Dong et al. [4], we modeled mutual exclusion of facts to improve the quality of the RE results. If relations of type '*player A* def. *player B*' and '*player B* def. *player A*' were both extracted, only the result of the most frequent relation is taken (both are taken in case they have the same frequency). Figure 2 shows the F_1 measure for each tournament when the results were filtered on mutual exclusion. The F_1 measure in general goes up, since the precision rises (the results are more likely to be correct). Good to note that the recall dropped 0.016 at WI and 0.031 at RG. The F_1 measure of WI increases 0.056 (10.8%) and for RG increases 0.037 (6.4%).

Modeling mutual exclusion works as at each tournament players cannot meet more than once. However, throughout their career they might meet multiple times. Therefore future research should be done to take into account the time constraints.

3.2 Fact repetition

The noisiness and the informal language widely used in tweets increase the uncertainties in the extracted information. First, the limited length of characters forces authors to use abbreviations which leads to the lack of context [12]. By limiting the dataset to only tweets which have a certain hashtag or keyword, the context can easily be established. Moreover, different authors have different writing styles. If a conservative RE method is used, this would lead to a low recall, since it is difficult to make a regular expression

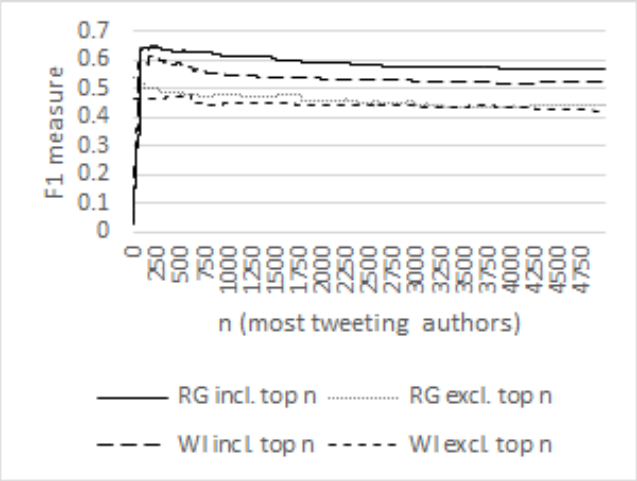

Figure 3: F_1 measure for tweets of top n authors

which matches all writing styles. A less conservative method could lead to a lower precision though, since it would extract more false positives. For example, our regex extracts the relation '*Djokovic* def. *Wawrinka*' from the text '*Djokovic* has been beaten by *Wawrinka*'. Therefore it is useful to check whether there's factors that can lead to more precision of the RE.

Laek and Vojt [10], Ji et al. [9], and Corney et al. [2] mentioned that the repetition of facts could give more confidence in the results. We check whether filtering the results on only those mentioned n times or more, leads to a better precision. Figure 2 shows the influence of different thresholds on the F_1 measure.

As the threshold is getting higher, the precision increases while the recall drops. This is because some results of some matches at the early rounds are tweeted few times. Hence, throwing their extracted results away leads to a lower recall. The F_1 measure gives the combination of precision and recall. Figure 2 shows that for RG the optimal threshold is 2 (an increase of 0.032 or 5.1%), and the F_1 measure drops quickly afterwards. At WI the peak is at a threshold of 3 (an increase of 0.048 or 8.4%), and stays a bit constant before it drops. This difference shows that the best threshold differs in every case. Future research needs to be done to find an automatic way of finding the optimal threshold and to check whether there is a correlation between the size of the corpus and the threshold (the WI-dataset is larger and also has a higher optimal threshold).

3.3 Trusted authors

Since no one verifies whether what is tweeted is true, it is difficult to say whether an author is trustworthy. Therefore, we did some analysis to check whether the most-tweeting accounts are more trustworthy than others. We also want to check whether some specific accounts (of journalists for example) provide true information. This analysis was done disregarding the mutual exclusion rule since it could bias the results (Figure 2 shows that with a lower threshold we consider more relations and thus the improvement resulted from using the mutual exclusion is more significant).

For this analysis, the tweets of the top n most-tweeting authors were considered. Analysis is done to check the effect of the contributions of the top authors if only their tweets are used for extraction or completely excluded from the dataset. Figure 3 shows the F_1 measures for each n. By considering only the tweets of the top n authors, the best results (the peak) are achieved at n=168 for WI and n=187 for RG. At this peak, the results of WI are improved by 0.15 (33%) and results for RG are improved by 0.11 (21%). After this peak, the F_1 results considering the n authors slowly goes down until it becomes similar to the value of including all authors (the figure has been limited to n=5000 but there are many more authors). It is good to note that the line excluding the tweets of the top n authors, only decreases gradually. This means that the results of analyzing the tweets of a large crowd are barely influenced by the tweets of a small group of authors (wisdom of crowd).

An analysis of the individual authors shows large differences within the top n authors as well, some giving a lot of extracted results with high precision and some others almost none. Our analysis of the best 15 authors (achieving the highest F_1) showed that 1 author uses a standard format (most probably a robot account). Beside this, 5 authors are, in most of the times, re-tweeting others and 6 authors seem to be regular people tweeting out of interest (not a sports professional or journalist). Since there is a substantial overlap in the authors of the two datasets, further research could be done to find out whether the existence of trustworthy authors can be generalized to Twitter as a whole. Furthermore, it is also useful to research whether the most trustworthy authors have certain characteristics in common such as having a high number of followers or having official accounts.

3.4 Semantic web

Our proposed extraction approach uses the semantic web for the task of NEL. The semantic web has the advantage that it has a lot of information available to aid in IE, like in this case the (nick)names of all players. A problem is that it is difficult to analyze the truth of these facts. Dong et al. [4] did some analysis on the amount of truthful facts in different semantic web datasets, but does not show the trustworthiness of each source. It is interesting to see the effect of different sources on IE. It might be that certain data sources lead to better results. In our case, DBPedia and Yago are analyzed to have an initial idea of the influence of the semantic web on IE.

To find the (nick)names, first the list of participants had to be linked to entries in the semantic web. Not all players could be found (no qualitative analysis was made whether the correct entries were matched, this is left to the analysis of the results extraction). In DBPedia, 246 out of 256 players playing both the men and women tournaments were found at WI and 240 out of 256 were found at RG, whereas using Yago all players at WI could be found and 253 out of 256 players at RG. Figure 4 shows the consequences for the RE. Remarkable is that the F_1 measure is higher for DBPedia because both precision and recall for the DBPedia-extractions are higher, whereas in Yago more player names were linked (quick analysis of the links showed that correct links were established).

Comparing the results automatically is only possible if DBPedia and Yago are linked. Future research will be needed to find out the origin of the differences in the results. Using multiple sources like DBPedia and Yago together could lead to even better results. The results show that different semantic web sources lead to some degree of uncertainty in the results.

3.5 Combination

The previous experiments show that the uncertainties in the relation extraction could be reduced by using domain knowledge, ensuring a number of repetitions, taking into account the most-tweeting authors and using a specific knowledge base. A test has been run to find the best combination of these factors. Mutual exclusion seems to be beneficial in all cases. When taking only the most-tweeting authors (n), the best results were achieved at a threshold of 1 for the minimum number of repetitions (T). Table 2

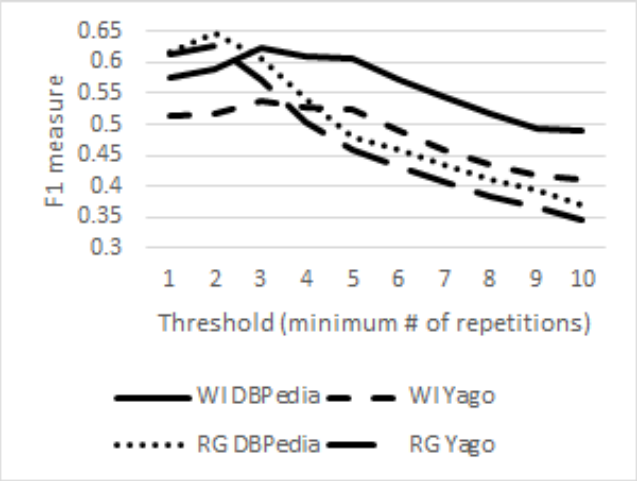

Figure 4: F_1 measures for DBPedia and Yago

Table 2: Best results for the combination

	n	T	F_1	Increase
Wimbledon	169	1	0.650	0.130 (25.0%)
Roland Garros	187	1	0.671	0.092 (15.9%)

shows the best combination found.

4. CONCLUSION & FUTURE WORK

In this paper, we investigated the uncertainties involved in the relation extraction process from a highly noisy data source (Twitter) with an application to extract games' results for two Grand Slam tennis tournaments. We used a modular approach of extraction and proposed different approaches to reduce the uncertainties involved. Our proposed approaches are to use domain knowledge, fact/relation repetition, wisdom of crowd, and choosing the right semantic web source. The combination of the proposed methods leads to an improvement of up to 25% in the extraction F_1 results.

In future research, more statistical analysis has to be done. We need to check the validity of the proposed methods on other domains. We also want to take time constraints into account. Furthermore, we plan to do some correlation analysis to automatically find the optimal number of repetitions that leads to the best balance between precision and recall. Corresponding characteristics for trustworthy authors can also be researched in more detail. Last, more research on the trustworthiness of semantic web datasets is needed and of the causes of the different extraction results.

5. REFERENCES

[1] M. Adedoyin-Olowe, M. Gaber, C. Dancausa, and F. Stahl. Extraction of unexpected rules from twitter hashtags and its application to sport events. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, pages 207–212, Dec 2014.

[2] D. Corney, C. Martin, and A. GÃűker. Spot the ball: Detecting sports events on twitter. In M. de Rijke, T. Kenter, A. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann, editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 449–454. Springer International Publishing, 2014.

[3] S. Dennis. An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5206–5213, 2004.

[4] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA, 2014. ACM.

[5] 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*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[6] M. B. Habib and M. van Keulen. Unsupervised improvement of named entity extraction in short informal context using disambiguation clues. In *SWAIE 2012 : Semantic web and information extraction*, volume 925 of *CEUR workshop proceedings*, pages 1–10. CEUR-WS.org, October 2012.

[7] M. B. Habib, M. van Keulen, and Z. Zhu. Concept extraction challenge: University of twente at #msm2013. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts', Rio de Janeiro, Brazil*, volume 1019 of *CEUR Workshop Proceedings*, pages 17–20, Aachen, Germany, May 2013. CEUR.

[8] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194(0):28 – 61, 2013. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

[9] H. Ji, H. Deng, and J. Han. Uncertainty reduction for knowledge discovery and information extraction on the world wide web. *Proceedings of the IEEE*, 100(9):2658–2674, Sept 2012.

[10] I. LaÅąek and P. VojtÃąÅą. Various approaches to text representation for named entity disambiguation. *International Journal of Web Information Systems*, 9(3):242–259, 2013.

[11] X. Liu, M. Zhou, F. Wei, Z. Fu, and X. Zhou. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 526–535, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[12] 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*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[13] I. Twitter. About twitter, 2015.

[14] G. Van Oorschot, M. Van Erp, and C. Dijkshoorn. Automatic extraction of soccer game events from twitter. volume 902, pages 21–30, 2012.

[15] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proceedings of the 3rd International Semantic Search Workshop*, SEMSEARCH '10, pages 6:1–6:10, New York, NY, USA, 2010. ACM.