SANE: System for Fine Grained Named Entity Typing on Textual Data

Anurag Lal
Indian Institute of Technology (BHU)
Varanasi, India–221005
anurag.lal.cse13@iitbhu.ac.in

Apoorve Tomer
Indian Institute of Technology (BHU)
Varanasi, India–221005
apoorvet.rs.cse16@iitbhu.ac.in

C. Ravindranath Chowdary
Member, ACM
Indian Institute of Technology (BHU)
Varanasi, India–221005
rchowdary.cse@iitbhu.ac.in

ABSTRACT

Assignment of fine-grained types to named entities is gaining popularity as one of the major Information Extraction tasks due to its applications in several areas of Natural Language Processing. Existing systems use huge knowledge bases to improve the accuracy of the fine-grained types. We designed and developed SANE, a system that uses Wikipedia categories to fine grain the type of the named entities recognized in the textual data.

The main contribution of this work is building a named entity typing system without the use of knowledge bases. Through our experiments, 1) we establish the usefulness of Wikipedia categories to Named Entity Typing and 2) we show that SANE's performance is on par with the state-of-the-art.

CCS Concepts

•**Information systems** → **Content analysis and feature selection; Information extraction;** *Entity resolution;*

Keywords

Named Entity Typing; Fined-grained; Wikipedia

1. INTRODUCTION

Named Entity Recognition (NER) is the task of identifying and classifying named entities in a sentence into predefined categories such as names of persons (PER), organizations (ORG), locations (LOC) or miscellaneous (MISC). We developed a system that exploits Wikipedia categorization to fine-grain this classification process (Named Entity Typing) of the Stanford NER.

In Named Entity Typing (NET), we associate semantic types of interest with a given entity name. For instance, given "Sachin plays cricket", our objective is to conclude that "Sachin" is a *cricketer* or *sportsperson* (NET) and a *person* (NER). In the attempt for a finer granularity, complex knowledge bases (KBs) like YAGO [9], DBPedia [1], etc. have been used. These KBs use complex algorithms to populate and organize entities into semantic categories. However, the rate of generation, modification and termination of entities is very high. This accounts for the delay in the incorporation of such emerging entities in the knowledge bases, which makes knowledge bases inherently incomplete.

Due to encyclopedic nature of Wikipedia, most of the articles are about named entities. The advantage with Wikipedia is that articles are frequently added or updated, so the entity set is up to date. Wikipedia categories are used to group similar pages based on their content. They are implemented by a MediaWiki[1] feature which adds any page with a text like [[$Category : ABC$]] in its wiki markup, to the category with name ABC. Categories can be found at the bottom of an article page. Category name links to a category page (a page in the category namespace) which lists the articles that have been added to that category.

KBs are huge in size and they need periodic human curating to maintain their consistency [3]. This motivated us to develop SANE which uses Wikipedia to assign fine-grained types to named entities.

2. BRIEF DESCRIPTION

In this section, we briefly describe the procedure that SANE uses for NET.

In the first phase, SANE looks for a set of patterns in the input sentence that explicitly refer to named entities. This is based on previous work by Hearst [5]. For example, if the sentence contains "Sachin Tendulkar, the cricketer", then the entity "Sachin Tendulkar" has the type "cricketer" explicitly mentioned. In this phase, such explicit types are extracted. We associate this type with the entity and directly go to the type selection phase.

We have used only four patterns that have high precision and do not lead to erroneous extractions. These patterns were chosen empirically. These patterns are listed in Table 1.

If explicit pattern-based extraction fails, then SANE falls back to the lookup-based extraction. In this phase, SANE

[1] `https://mediawiki.org/wiki/MediaWiki`

WWW'17 Companion, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054724

Figure 1: Overview of SANE

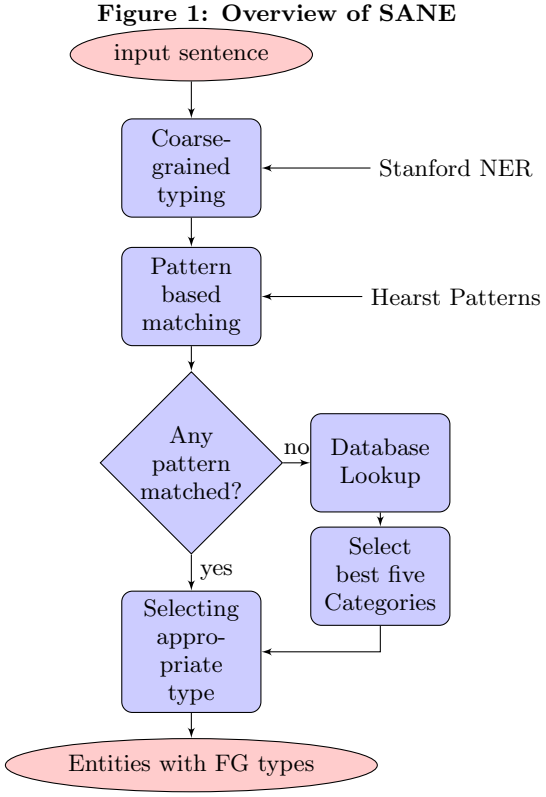

Table 1: Patterns used in SANE [2]

Pattern	Example
Hearst	[Sachin] and other {cricketers}
Apposition	[Sachin], the {cricketer}
Copular	[Sachin] is a {cricketer}
Noun modifier	{Cricketer} [Sachin]

exploits Wikipedia categorization by performing lookup in a database containing article names and corresponding categories. We generated this database from the English version of Wikipedia XML dump at the time January, 2014 [2]. For *locations*, we created a database from Geonames[3], consisting of 202 countries (including location aliases), 3893 states, and 41023 cities. In a previous version of SANE, we used Wikipedia categories for $LOCATION$ also but the quality of results was not good. After SANE performs the database lookup, in which the article titles are matched with the named entity and corresponding categories are added to a category list, it selects the best five categories based on our selection model. Also, SANE uses Word2Vec [6] in the selection model.

In a subsequent type selection phase, the explicit type from the pattern-based extraction phase or the categories selected in the lookup-based extraction phase are mapped to appropriate WordNet [7] types. Once the WordNet types are found, SANE tags the named entity with the most appropriate type which is also selected based on our selection criteria.

[2]`https://dumps.wikimedia.org/enwiki/`
[3]`http://geonames.org/`

3. EXPERIMENTAL SETUP AND RESULTS

We conducted experiments, using Twitter dataset to compare SANE with a state-of-the-art system FINET [2]. FINET is a system for fine-grained typing of named entities in context. It makes use of multiple extractors for extracting both explicit and implicit types and then selects appropriate type in a subsequent type-selection phase. The type selection phase uses principles of Word Sense Disambiguation that are adjusted for fine-grained NER. It also leverages WordNet as its type system.

Data: We extracted 1000 tweets using the Twitter API. We formatted these tweets (like removing hashtags, etc.) in order to simplify the tagging. We specifically chose Twitter dataset as tweets usually contain entities that have Wikipedia articles. We selected only those tweets that contained at least one entity according to Stanford NER. We compared SANE and FINET on this dataset.

System: SANE uses the Stanford NER 3-class classifier to identify named entities with their coarse-grained category types in the tweets. We have classified the category types into coarse-grained (CG) and fine-grained (FG). SANE processes the identified entities using the procedure outlined in Section 2. The CG category system consists of three categories – PERSON, LOCATION, and ORGANIZATION while FG category system consists of hyponyms of $< person-1 >$ and $< imaginary\_being-1 >$ as $PERSON$, hyponyms of $< organization-1 >$ as $ORGANIZATION$ and hyponyms of $< location-1 >$ as $LOCATION$. FINET's FG type system consists of 200 WordNet types that are included in Pearl [8]. FINET also generates super fine-grained (SFG) types that we are ignoring for the purpose of comparison.

Table 2: Summary of Results

System	Total Entities	Correct Types	Precision
SANE CG	1640	1588	96.83
FINET CG	1638	1587	96.88
SANE FG	1284	1012	78.82
FINET FG	1210	966	79.80

Table 3: SANE Extractor-wise performance

Extractor	Entities	Precision
Pattern-based	22	90.91
Lookup-based	990	78.45

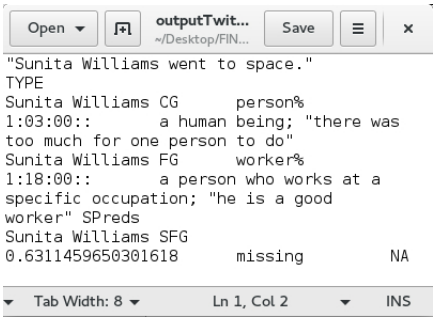

Figure 2: FINET output

Also, the precision of FINET for the SFG types is less than that for the FG types.

Labelling process:Two independent annotators label the results. We considered a category type to be correct iff it was labeled correctly by both the annotators.

3.1 Result Analysis

The performance of FINET and SANE is given in Table 2. For each system, the table shows the total number of entities for which category types were extracted, the number of correct category types and the precision for both CG and FG types. The Cohen's kappa measure[4] is 0.72 for FINET and 0.86 for SANE indicating high inter-annotator agreement.

SANE assigns types to more entities as compared to FINET and has comparable precision. The slight dip in precision can be attributed to the fact that SANE does not depend on supervised or unsupervised learning and has a very simple design whereas FINET has a complex design.

As can be seen from the Figure 2, FINET tags "Sunita Williams" as *worker*, however from Figure 3 SANE tags "Sunita Williams" as *American*. SANE associates a finer type to person "Sunita Williams". For some cases, SANE is not able to identify a correct FG, for e.g. in Figure 4 SANE

[4]The Cohen's kappa measure is the overall score that includes both CG and FG types

Figure 3: SANE output1

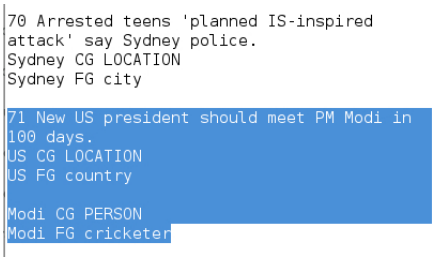

Figure 4: SANE output2

tags "Modi" as a *cricketer* where as the presence of "PM" changes the context to the *head_of_state*.

Table 3 shows the extractor-wise performance. We observe that the pattern-based extractor has a lower recall but higher precision, whereas, lookup based extractor has a higher recall but lower precision.

The incompleteness of Wikipedia is a source of error. Some articles don't have any categories while some other articles have tens of categories making it difficult to select a category that fits into the context. Context can be taken into account in a better way by considering not only the words present in the sentence but also their lexical expansions. The category selection phase selects inappropriate categories for some entities as word2vec does not work very well with proper nouns. Due to the presence of incomplete names of entities, SANE sometimes selects the wrong context.

3.2 Values and Contributions

As compared to existing *NET* systems like FINET [2], Hyena [10], our system SANE does not depend on knowledge bases for type extraction. We demonstrate in this paper that category labels extracted from categories of a Wikipedia article are useful to improve the granularity of NER. For example, "Banaras Hindu University" has the article associated with categories such as "Universities and colleges in Uttar Pradesh", "Educational Institutions established in 1916" and "Indian academics". These categories seem to be extremely useful for NER. We use such category labels to fine-grain the results of Stanford NER [4]. In our experiments, we used Twitter dataset to demonstrate that we can improve the granularity of the results of Stanford NER using Wikipedia categorization. To sum up the contributions:

1. We demonstrate the usefulness of Wikipedia categories for named entity typing,

2. Our system incorporates only pattern-based and lookup-based extractors, which makes it simple as compared to the state-of-art systems.

4. REQUIREMENTS FOR THE DEMO

We will present the demo using sample tweets from our dataset from which named entities will be identified and

categorized. SANE is developed in python 3 and integrated with Stanford NER which is built in java. So we would require python 3.0, JDK 1.8 and Linux OS on a system with atleast 16GB RAM. A video demonstration is published on Youtube[5].

5. ABOUT THE AUTHORS

1. **Anurag Lal** is an undergraduate student at the Indian Institute of Technology (Banaras Hindu University), Varanasi. He has particular interest in the field of Information Retrieval.

2. **Apoorve Tomar** is a graduate student at the Indian Institute of Technology (Banaras Hindu University), Varanasi.

3. **Dr C. Ravindranath Chowdary** is an Assistant Professor in the Department of Computer Science and Engineering at Indian Institute of Technology (Banaras Hindu University), Varanasi. He pursued his PhD from Indian Institute of Technology Madras in 2009. His research areas include Information Extraction and Web Mining.

[5]`https://www.youtube.com/watch?v=_mHtfsJUIAo`

6. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[2] L. Del Corro, A. Abujabal, R. Gemulla, and G. Weikum. Finet: Context-aware fine-grained named entity typing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 868–878, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[3] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. Building, maintaining, and using knowledge bases: A report from the trenches. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 1209–1220, New York, NY, USA, 2013. ACM.

[4] J. R. Finkel, T. Grenager, and C. D. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In T. A. for Computer Linguistics, editor, *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, 2005.

[5] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[7] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[8] N. Nakashole, T. Tylenda, and G. Weikum. Fine-grained semantic typing of emerging entities. In *ACL (1)*, pages 1488–1497. The Association for Computer Linguistics, 2013.

[9] F. M. Suchanek, G. Kasneci, and G. Weikum. 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, 2007. ACM.

[10] M. A. Yosef, S. Bauer, J. Hoffart, M. Spaniol, and G. Weikum. Hyena-live: Fine-grained online entity type classification from natural-language text. In *ACL (Conference System Demonstrations)*, pages 133–138. The Association for Computer Linguistics, 2013.