Improving Information Extraction from Wikipedia Texts using Basic English

Teresa Rodríguez-Ferreira, Adrián Rabadán, Raquel Hervás, Alberto Díaz
Facultad de Informática
Universidad Complutense de Madrid
teresaro@ucm.es, arabadan@ucm.es, raquelhb@fdi.ucm.es, albertodiaz@fdi.ucm.es

Abstract

The aim of this paper is to study the effect that the use of Basic English versus common English has on information extraction from online resources. The amount of online information available to the public grows exponentially, and is potentially an excellent resource for information extraction. The problem is that this information often comes in an unstructured format, such as plain text. In order to retrieve knowledge from this type of text, it must first be analysed to find the relevant details, and the nature of the language used can greatly impact the quality of the extracted information. In this paper, we compare triplets that represent definitions or properties of concepts obtained from three online collaborative resources (English Wikipedia, Simple English Wikipedia and Simple English Wiktionary) and study the differences in the results when Basic English is used instead of common English. The results show that resources written in Basic English produce less quantity of triplets, but with higher quality.

Keywords: Information Extraction, Triplets, Basic English

1. Introduction

Although software applications could theoretically benefit from the huge amount of information in the Web, they usually face the problem of this information appearing in the form of unstructured data like plain text. The possibility of automatically extracting the knowledge underlying this plain text is therefore becoming increasingly important.

Information Extraction (IE) is the process of automatically extracting structured data from unstructured texts. There are different ways to represent data extracted from text, such as in the form of graphs or by using triplets in the form *(concept$_1$, verb, concept$_2$)* to express relations between concepts extracted from the text. Although there are many IE approaches, in this paper we are only interested in unsupervised techniques that are able to extract information from plain text. For this kind of technique, the characteristics of the source text from which the information is going to be extracted play an important role in the obtained results.

In this paper we will evaluate whether the use of Basic English instead of common English leads to the extraction of more accurate data by implementing an experiment that compares triplets extracted from the English Wikipedia[1], Simple English Wikipedia[2] and Simple English Wiktionary[3] (from now on referred to as Simple Wikipedia and Simple Wiktionary). Basic English is a simplification of the English Language created by Ogden (1930) which defends that full communication can be achieved by using only 850 English words. In addition to using Basic English, Simple Wikipedia and Simple Wiktionary also ask users to write in shorter sentences, use active voice over passive voice and provide guidelines to help users write sentences with simple structures.

The triplets used will represent definitions and properties, concepts that establish a unidirectional IS_A or IS relation with certain other concepts. Even though these two relations are different, they can both be used to define a concept, so they have not been considered separately in the final results. This type of output will be easily computable by machines and can be used to establish new relations between concepts. This can be achieved, for instance, by connecting triplets in which the second concept is the same as the first concept of the other triplet.

The paper will address questions such as:

- Are triplets obtained from text written in Basic English more useful?
- How does information obtained from dictionaries compare to information obtained from encyclopedias?

The goal of this work is not to provide a new IE technique that improves previous work results, but to demonstrate that texts written using simplified vocabulary and grammar will lead to better triplet extraction.

In Section 2 we discuss previous work that is relevant to the field of Information Extraction. In Section 3 we describe the sources used and the results we expect to obtain from them, and we cover implementation details. In Section 4 we explain the evaluation criteria for the quality of the triplets obtained, we present the final results and we cover the issues encountered during this research. Section 5 is a discussion of the results. Finally, Section 6 describes future work that will improve the triplet extraction system.

2. Related work

Information Extraction (IE), the process of automatically extracting structured information from unstructured texts, has progressed substantially over the last few decades (Etzioni et al., 2008). Although the ambiguous nature of plain text makes the task an arduous one, it is possible to find many systems that have obtained quite good results. TextRunner (Yates et al., 2007), one of the pioneers in Open Information Extraction (OIE), is able to obtain high-quality information from text in a scalable and general manner.

[1] http://www.wikipedia.org
[2] http://simple.wikipedia.org
[3] http://simple.wiktionary.org

Rusu et al. (2007) present an approach to extracting triplets from sentences by relying on well known syntactic parsers for English.
Wikipedia is considered an excellent source of texts for IE systems due to its broad variety of topics and advantageous characteristics such as the quality of the texts and their internal structure. Therefore there are some IE systems that work with Wikipedia texts and/or their structured metadata, like Wanderlust (Akbik and Bross, 2009) or WOE (Wikipedia-based Open Extractor) (Wu and Weld, 2010). Weld et al. (2009) restrict their process to infoboxes, tabular summaries of an article's salient details which are included in a number of Wikipedia pages. Wanderlust (Akbik and Bross, 2009) is an algorithm that automatically extracts semantic relations from natural language text. The procedure uses deep linguistic patterns that are defined over the dependency grammar of sentences. Due to its linguistic nature, the method performs in an unsupervised fashion and is not restricted to any specific type of semantic relation. The applicability of the algorithm is tested using the English Wikipedia corpus. WOE (Wikipedia-based Open Extractor) (Wu and Weld, 2010) is a system capable of using knowledge extracted from a heuristic match between Wikipedia infoboxes and corresponding text. In particular, Krawczyk et al. (2015) present a method of acquiring new ConceptNet triplets automatically extracted from Japanese Wikipedia XML dump files. In order to check the validity of their method, they used human annotators to evaluate the quality of the obtained triplets.

3. Using Basic English for improving Information Extraction from texts

Our goal is to extract triplets which represent definitions or properties of a given concept established by a unidirectional IS_A or IS relation. Many other relations can be considered, but they are out of the scope of this experiment.

3.1. Textual knowledge sources

The sources where the triplets are extracted from must contain definitions and properties of concepts. The most appropriate resources for this purpose are dictionaries and encyclopedias. Dictionaries provide succinct definitions and a brief and usually more technical overview of the concept's most salient properties. Encyclopedias, on the other hand, contain more general information and in greater quantity.
We have chosen to use Wikipedia, Simple Wikipedia and Simple Wiktionary as sources for Information Extraction. All three are free-access and free-content collaborative Internet encyclopedias or dictionaries. This type of resource is fast-growing, with content created by users from all over the world (refer to Table 1).
Wikipedia is ranked as one of the top ten most popular websites at the time this article is written, so it provides a rich source of general reference information for this type of work. One of the main concerns when using a free-content resource is the quality of its content and language. Since we are not going to attempt to extract complex details of the concepts, the accuracy of these sources does not pose an impediment, because their general definitions tend to be correct. On the other hand, the structure of the text can be problematic when parsing the information. A simple grammatical error or an incorrectly structured sentence may lead to no triplets being extracted, or to triplets containing properties which are not definitions of the concept. This type of error is more likely to occur in sources where articles are longer and more complex.
Below is an example of a fragment of text extracted from the same article for each of the different sources:

1. Wikipedia: "Chocolate is a typically sweet, usually brown, food preparation of Theobroma cacao seeds, roasted and ground, often flavored, as with vanilla. It is made in the form of a liquid, paste, or in a block, or used as a flavoring ingredient in other foods."

2. Simple Wikipedia: "Chocolate is a food made from the seeds of a cacao tree. It is used in many desserts like pudding, cakes, candy, and ice cream. It can be a solid form like a candy bar or it can be in a liquid form like hot chocolate."

3. Simple Wiktionary: "Chocolate is a candy made from cacao beans and often used to flavour other foods such as cakes and cookies. A chocolate is an individual candy that is made of or covered in chocolate. Chocolate is a dark brown colour."

3.2. Triplet extraction

In order to extract relevant semantic information from the text, it must first go through a process of morphological analysis and dependency parsing. The analyser used was Freeling 2.2 (Carreras et al., 2004), an open source language analysis tool suite that supports several languages, including English.
The information for each specified concept was obtained from the corresponding web page from each source. For example, for the concept *pinneaple* and the source Simple Wikipedia the wiki page used was *https://simple.wikipedia.org/wiki/Pineapple*. This information was parsed into plain text, and then morphologically analysed using Freeling 2.2 (Carreras et al., 2004). This was in turn used as input for the dependency parsing, producing a final output of a tree containing all the semantic information. After this, the objective was to extract only IS_A or IS relations from the texts, so only sentences which had as their root any form of the verb "to be" were considered. Assertions that make use of a form other than the present tense were taken into consideration because texts referring to historic events or characters may use the past tense. Once the relevant sentences had been collected, the next step was to find the ones referring to the specified concept. Since the aim is to extract IS_A or IS relations, the third element of the triplets is always a definition or a property of the first element, so the triplets follow this structure: *concept - verb - property*.
In order to obtain definitions of the concept or related information from the text, the object of the chosen sentences has been studied. There are three possible scenarios depending on the root of the object (refer to Table 2):

1. When the root of the object is a noun, it is considered as a possible definition of the concept. For instance

-	English Wikipedia	Simple Wikipedia	Simple Wiktionary
Articles	4,977,081	115,138	24,309
Users	26,395,232	470,736	14,981
Articles per user	0.19	0.24	1.62

Table 1: Usage statistics of the used resources

in the sentence "A pineapple is a fruit", the object is "a fruit" and its root is "fruit", which is a noun, so it is saved in a triplet (pineapple - be - fruit). This represents an IS_A relation.

2. If the noun has any modifiers which are adjectives, they are also selected as possible information related to the concept. For instance in the phrase: "Chocolate is a dark brown colour", the root of the object ("colour") has two modifiers, "dark" and "brown", so aside from the triplet that represents an IS_A relation (chocolate - be - colour), both adjectives are stored in additional triplets (chocolate - be - dark, chocolate - be - brown). This type of information represents a property of the concept, an IS relation.

3. If the root of the object is the conjunction "and" or "or" instead of a noun, its children are searched for nouns and adjectives much like in the previous case, for example in the sentence "Battle Royale is a novel and a film" (Battle_Royale - be - novel, Battle_Royale - be - film). This represents an IS_A relation when the child is a noun or an IS relation when it is an adjective.

As an example, we can observe the differences between the properties extracted for the concept "wine":

- From Wikipedia, the extracted properties for the triplets were *cabernet_sauvignon*, *gamay*, *merlot*, *part*, *tradition* and *red*.
- From Simple Wikipedia, the properties were *drink*, *alcoholic* and *popular*.
- From Simple Wiktionary, only one property was extracted: *drink*.

4. Evaluation

The evaluation criteria used to verify the quality of the extracted triplets is similar to the one used by Krawczyk et al. (2015). Every triplet generated for each concept is assigned a value based on how strongly related its property is to the concept and how well it respects the relation. The possible values are 1, 0.5 and 0.

- Triplets get the highest score when they correctly represent an IS_A or IS relation in which the property defines or is very strongly related to the concept. For instance the triplet *car - be - vehicle* would be considered a good triplet and it would be assigned 1 point.
- Mediocre triplets are assigned 0.5 points, when the property is a less accurate or informative definition of the concept, or when it represents a feature or quality of the concept. Note that the IS_A or IS relation must still be respected. A triplet such as *book - be - product* would have a score of 0.5 points.
- Triplets with properties which are related to the concept but do not respect the relation (for example *moon - be - crater*) or which are unrelated to the concept (*chocolate - be - iron*) are considered bad triplets and receive the lowest score (0).

The evaluation so far has been performed manually by four human annotators. The triplets generated for this evaluation were divided into four groups, where each annotator evaluated two groups and each triplet was evaluated by two annotators. The final statistics were obtained by using the average of the score given by all of the annotators, following an inter-annotator agreement using a popular metric, Fleiss Kappa (Fleiss, 1981). This allows us to know the degree of agreement between the annotators.

4.1. Results

A total of 62 concepts were randomly chosen as input (e.g.: pinneaple, chocolate, Battle Royale...), 49 of which generated triplets for at least one of the knowledge sources. The absence of triplets for some concepts is due to texts with sentences defining the concept which do not match the required pattern accepted by the extractor. Both common nouns (water, yellow, chair...) and proper nouns (New York, Bruce Willis, Final Fantasy...) were used as input, and the latter produced less triplets (7 of the 13 concepts that did not generate any triplets were proper nouns). A total of 604 triplets were examined (428 from Wikipedia, 124 from Simple Wikipedia and 52 from Simple Wiktionary).
The results reflected in Table 3 show that sources with a large amount of content produce triplets for more concepts, as was expected. Consequently, Wikipedia is the source that offers the most good triplets (those assigned 1 point), followed by Simple Wikipedia and Simple Wiktionary. Note however that it also produces more mediocre triplets (0.5 points) and many more bad triplets (0 points) than the others. Even though the quantity of the triplets generated for sources using Basic English is compromised, their quality is much higher. Less than a third of the triplets extracted from Wikipedia can be considered good, and less than 10% are mediocre. This means that around 64% are bad triplets, representing information that is not related to the specified concepts or that does not represent an IS_A or IS relation. Triplets extracted from Simple Wikipedia behave better, more than 40% of them are good, and less than half are bad. As shown in Table 3, the degree of agreement between triplets extracted from Wikipedia and Simple Wikipedia is more or less the same. The Kappa score for Simple Wiktionary is better and shows that the annotators

Sentence	Freeling V2.2 tree	Triplets
A pineapple is a fruit	claus/top/(is be VBZ -) [n-chunk/ncsubj/(Pineapple pineapple NN -) sn-chunk/dobj/(fruit fruit NN -) [DT/det/(a a DT -)]]	Pineapple - be - fruit
Chocolate is a dark brown colour	claus/top/(is be VBZ -) [n-chunk/ncsubj/(Chocolate chocolate NN -) sn-chunk/dobj/(colour colour NN -) [DT/det/(a a DT -) attrib/ncmod/(dark dark JJ -) attrib/ncmod/(brown brown JJ -)]]	Chocolate - be - dark Chocolate - be - brown Chocolate - be - colour
Battle Royale is a novel and a film	claus/top/(is be VBZ -) [n-chunk/ncsubj/(Royale royale NNP -) [NN/ncmod/(Battle battle NN -)] sn-coor/dobj/(and and CC -) [sn-chunk/conj/(novel novel NN -) [DT/det/(a a DT -)] sn-chunk/conj/(film film NN -) [DT/det/(a a DT -)]]]	Battle_Royale - be - novel Battle_Royale - be - film

Table 2: Triplet extraction scenarios

agree more on the quality of these triplets. Since the average score is higher for this source, this proves that triplets extracted from Simple Wiktionary have an overall better quality than the others.

The amount of concepts that generated triplets was similar for both Wikipedia and Simple Wikipedia, which means that the main difference between them was the content of the text. This proves that text expressed in Basic English yields more useful definitions for concepts than text written in common English.

Finally, the best results are achieved in Simple Wiktionary. Around 55% of the generated triplets are good definitions of the concepts, slightly less than 20% are mediocre, and less than a third of the triplets are bad. This seems to indicate that sources which contain less detailed and more specific content tend to result in higher quality triplets. Dictionaries are ideal, since they strive to define concepts briefly and do not offer additional background information.

	Wikipedia	Simple Wikipedia	Simple Wiktionary
Concepts with triplets	46 (74.19%)	40 (64.52%)	26 (41.94%)
Triplets	428	124	52
Good triplets	119 (27.8%)	54.5 (43.95%)	28.5 (54.81%)
Mediocre triplets	36.5 (8.53%)	12.5 (10.08%)	9 (17.31%)
Bad triplets	272.5 (63.67%)	57 (45.97%)	14.5 (27.88%)
Average score	0.32	0.49	0.63
Inter-annotator agreement (kappa)	0.496	0.49	0.578

Table 3: Results from the evaluation

4.2. Detected errors in triplet extraction

The above method is relatively simple to understand and to implement, but it has a few disadvantages. When the text does not have any sentences that match the required pattern exactly, no triplets can be extracted. For instance, if a definition uses a verb other than "to be", but equivalent to it, the sentence will be ignored. The definition of "purple" extracted from the Wikipedia ("Purple is defined as a deep, rich shade between crimson and violet [...]") cannot be processed because "defined" is the main verb and "is" is an auxiliary verb. If the word "is" had been used by itself, the triplets *purple - be - shade, purple - be - deep* and *purple - be - rich* could have been extracted.

As explained above, when the object's root is a noun with an adjective that refers to it, both noun and adjective are

stored separately in different triplets. In some cases the concept's definition only makes sense when the adjective and noun are used together. For example, when defining a foot, the sentence "anatomical structure" was obtained. This makes sense as a combination, but a person would not usually describe a foot as a structure. When this situation arises, both words usually make sense separately as well as combined, but in some cases storing them separately renders one or both of them useless. The final decision was to keep the information separately in the triplets, ensuring that the results will be more easily computable, at the expense of having triplets which are more general and less precise.

Accepting any form of the verb "to be", including past tense, means that relevant information can be extracted from text regarding past events or historical characters. The problem is that this could also result in out of date information. For instance the sentence "In ancient times Germany was largely pagan" results in the triplet *Germany - be - pagan*. This is not true at present, and so this triplet is incorrect.

An interesting phenomenon that occurs is when providing examples of a concept. Sentences such as "Examples of [concept] are..." or "A type of [concept] could be..." match the pattern recognised by the triplet extractor, so the sentence "A popular toy of this type is the Teddy Bear" will result in the triplet *toy - be - teddy_bear*. This represents information that is related to the concept, but since it does not match the IS_A or IS relation, it cannot be considered correct.

Another problem presents itself in articles about people or characters. Sometimes they are referred to in different ways inside the article, for instance by their full name, just their first name, just their surname or even a nickname. When searching for "Bruce Willis" in the Wikipedia, he is referred to as "Walter Bruce Willis" and further ahead as just "Willis". In this case only the sentences that contain the concept written exactly as specified can be examined.

5. Conclusions

The results discussed in section 4.1. reveal that sources written in Basic English produce less quantity of triplets for a given concept than those written in English, but the triplets display much higher quality. Overall, the triplets extracted from Simple Wiktionary are twice as good as those extracted from Wikipedia. Generally, longer articles which tend to be more detailed and provide background information about the concept result in more incorrect triplets. This can be observed especially in articles concerning very general topics or articles on historical events and characters, for instance in the article regarding the Earth. For this reason, articles from Wikipedia, which are usually longer than those in Simple Wikipedia, produce more triplets per concept, but a large portion are incorrect. On the other hand, certain types of articles do not produce any triplets, especially articles regarding proper nouns (such as countries, cities, books, films, games or names of people). In our evaluation 15 concepts which are proper nouns were introduced, and roughly half of them (7) did not generate triplets for any of the sources.

The precise and succinct style of dictionaries seems more useful in the extraction of IS_A and IS relations between concepts and their properties. The triplets extracted from this type of source are also more easily evaluated by human annotators, since the information they contain is more objective. More research is needed, however, in order to correctly compare results extracted from encyclopedias against results extracted from dictionaries.

6. Future work

In this research, our goal was to compare triplets obtained from sources written in common English with those from sources written in Basic English. For this reason Wikipedia and Simple Wikipedia were the first two options to be considered. While analysing the results obtained, it seemed likely that Simple Wiktionary might be an even better source than Simple Wikipedia. This was on the grounds that aside from using Basic English and simpler sentence structure, its content is more precise and focuses solely on definitions, which was the goal of this study. We did not, however, evaluate results obtained from the English Wiktionary. It would be interesting to compare Simple Wiktionary against Wiktionary to examine the effect of IE from dictionaries written in Basic English, and to compare Wikipedia against Wiktionary to further observe the differences between data extracted from dictionaries and from encyclopedias. However, these resources could also be combined since the information contained in each one complements the others.

The extracted triplets follow a simple structure: *concept - verb - property*. In this work the verb used is always "to be", but this could be extended to also include relations such as HAS_A or RELATED_TO.

Having encountered the errors discussed in section 4.2., it would be useful to detect the patterns that lead to these errors and address them before saving the triplets. The matter of storing nouns and the adjectives that apply to them separately or together should also be explored further. When stored separately they lead to a larger amount of simpler triplets, but some information can be lost in the process, leaving either the noun or the adjective meaningless without its partner. Finally, the use of synonyms can aid in the recognition of additional triplets in the content. When searching for a concept, definitions that refer to it with a synonym (or a nickname or alternative name in the case of a person) are currently ignored. Using synonyms for common names, or alternative names found for people in sources such as DBpedia could produce richer results.

In order to reduce the time employed in the evaluation of the generated triplets, an automatic or semi-automatic criteria for evaluation should be implemented. By using existing triplets or relations similar to ours from sources such as ConceptNet, we could compare the results with others that are accepted as correct to automatically approve the common triplets.

7. Acknowledgements

This work is funded by ConCreTe. The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European

Commission, under FET grant number 611733.

8. References

Akbik, A. and Bross, J. (2009). Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In *Proceedings of the 2009 Semantic Search Workshop at the 18th International World Wide Web Conference*, pages 6–15, Madrid, Spain.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Commun. ACM*, 51(12):68–74, December.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. Wiley, New York, USA.

Krawczyk, M., Rzepka, R., and Araki, K. (2015). Extracting conceptnet knowledge triplets from japanese wikipedia. In *21st Annual Meeting of The Association for Natural Language Processing (NLP-2015)*, pages 1052–1055, Kyoto, Japan.

Ogden, C. K. (1930). *Basic English: A General Introduction with Rules and Grammar*. Paul Treber & Co., Ltd, London.

Rusu, D., Dali, L., Fortuna, B., Grobelnik, M., and Mladenic, D. (2007). Triplet extraction from sentences. In *Proceedings of the 10th International Multiconference Information Society*, pages 8–12.

Weld, D. S., Hoffmann, R., and Wu, F. (2009). Using wikipedia to bootstrap open information extraction. *SIGMOD Rec.*, 37(4):62–68, March.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., and Soderland, S. (2007). Textrunner: Open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, NAACL-Demonstrations '07, pages 25–26, Stroudsburg, PA, USA. Association for Computational Linguistics.