2P2-6

Commonsense and context: a novel approach for automatic extraction of generic statements

ダンコヴ　スヴェトスラヴ　　ジェプカ　ラファウ　　荒木　健次
DANKOV Svetoslav　　RZEPKA Rafal　　ARAKI Kenji

北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

In this paper we present a novel approach for automatically identifying common sense knowledge from unstructured text sources. We focus on the grammaticality of such statements as opposed to their semantic value. We identify such statements by evaluating the extent of their "genericity" and further refine our results by looking at other contextual clues extracted from the sentence.

1. Introduction

In order to create artificial intelligence systems that can relate to, make decisions about, and have a simple understanding of the global environment in which they operate, they need to be provided with a large source of basic knowledge about the world. Most artificial intelligence systems of our age are very domain specific, and thus are able to operate within a very confined set of parameters, lacking the general knowledge and reasoning shortcuts that common sense provides. With humans, this information is acquired naturally during our development stages, through both knowledge acquisition and reasoning based on what is already known. Computer agents, however, have no inherent mechanisms to acquire common sense knowledge or to derive inferences based on it. These mechanisms have to be supplied by the creators of the agent. Many projects are currently involved in manually providing such taxonomies but this process is costly and laborious. The World Wide Web, however, provides a ready source of common sense information that we can use. However, automatically identifying common sense in an unstructured text is a hard task as it is necessary to understand the general meaning of the text in order to do that. In this paper we will show that there exist syntactic and semantic clues that can successfully help us identify common sense statements.

2. Background

2.1 What is common sense

The term “common sense” is used to describe the collective shared experience of a particular culture or group of people. This experience may lie in any particular domain, be it social, economic, pragmatic, political, etc. This shared experience and the knowledge acquired from it is perceived to be universally true by the members of the particular culture. The term, unfortunately, coins a name for a phenomenon very difficult to quantify or describe in detail. The information considered common sense in any culture includes many different variations and most times overlaps with the term cultural (or personal) beliefs. What makes common sense difficult to work with is the fact that it does not simply represent information but the result of a reasoning process about some information. In this paper we attempt to look at the grammaticality of common sense expressions as opposed to the actual reasoning involved in formulating such a statement.

2.2 Previous research

There have been several attempts to collect common sense data. Two of the most prominent projects are the Open Mind Common Sense and the Cyc projects [Lenat 90]. However, both involve manual labor to collect the information. OMCS collects free form common sense statements fromvolunteers, while the Cyc project employs professionals to graft the structure of a domain and describe the common sense information involved in it. ConceptNet is a project based on the data already collected by OMCS, which provides a simple semantic structure of the collected statements in an attempt to make this data more accessible to researchers [Liu 04]. In addition to that the Cyc project has developed a reasoning language, CycL, used in making valid assertions and queries to their database.

Very few have attempted to develop automatic methods of collecting common sense statements with much success. The most notable attempt is also focus on identifying the common sense orientation of noun phrases only as opposed to looking at the sentence as a whole. Later in our paper, we will show that our results surpass those attempts.

2.3 Genericity

As a basis of our approach we employ the linguistic phenomena termed “genericity” and the syntactic structures used to represent it. A generic statement is defined as a *reference to a kind*, as shown in the sentence “The potato was first cultivated in South America”, where “the potato” is a kind-referring noun phrase. A different, but complimentary notion of generic statements is a proposition describing a kind of general property as opposed to a specific episode or isolated fact – as seen in the sentence “John smokes a cigarette after dinner.” Clearly, this second notion of “genericity” is a feature of the whole sentence as opposed to the kind-referring noun phrase in the first example, which is simply a feature of the noun phrase itself [Carlson 95]. For our purposes we use the kind-referring noun phrases as a starting point of our approach. It is important to point out that

Contact: Svetoslav Dankov, 北海道大学 情報科学研究科, 〒 060-0814, 札幌市北区北１４条西９丁目, dankov@media.eng.hokudai.ac.jp

one can view “generic” statements as a superset of common sense statements. Thus we need to look for further clues in order to refine our results [Carlson 82]. In this experiment we look at other syntactic contextual clues as well as additional semantic information.

3. Method

Our basic method begins at evaluating the genericity of a statement. First, sentences starting with a kind-referring noun phrase are selected as candidates. In the second step we look at the syntactic and semantic contexts of the candidates.

As syntactic context we consider the adjectives and adverbs found in the noun groups and verb group in the subject-verb-object relationship. For example, any sentence where either the noun group or the verb group has any of the following adjectives or adverbs – usual(ly), common(ly), frequent(ly), typical(ly), most(ly), every, all, most of, some of – will be selected as common sense candidate. In the cases when looking at the syntactic context fails, we use the WordNet semantic database to look at the semantic context of the sentence.

What we consider as semantic context is how frequently the subject and the verb of the sentence are actually used together. We select the subject noun and 3 of its synonyms (according to WordNet). We also look at the example set of the 3 most common uses of the verb of the sentence and we check if they contain as their subjects any of the nouns we selected in the previous step.

4. Evaluation

4.1 Corpus

For the purposes of our evaluation we are using the November 2006 snapshot of the XML Wikipedia article database and have selected 16,475 articles at random. We have selected only the textual parts of those articles, discarding titles and any irrelevant information. The articles are preprocessed with a tokenizer, sentence splitter, part of speech tagger and the SNoW shallow parser.

4.2 Evaluators

The extracted common sense statements were evaluated by two native speakers of English. The evaluators exhibited agreement of k = 0.701 during the evaluation, which shows a substantial agreement between them. The evaluators marked each statement as either being common sense (marked as “Yes” in Table 1), not common sense (marked as “No” in Table 1) or vague – if its generic meaning depended largely on the context of the text in which it appeared.

4.3 Results

Out of the 16,475 articles, our algorithm found 1,305 common sense candidate statements in 560 separate articles. This represents 3.4% coverage on the original set of articles. The results are summarized in Table 1, where the scores of the first evaluator are shown in the columns and those of the second evaluator – shown in the rows.

As we can see the number of statements on which both evaluators agree in their judgment is 1,124.

	Yes	No	Vague
Yes	**635**	45	35
No	50	**379**	11
Vague	24	16	**110**

Table 1: *Results of evaluation experiment*

Of the statements where both evaluators agreed, 56.5% were marked as common sense, 33.7% were marked as not common sense and 9.8% were marked as being too vague. The method described in [Suh 06] evaluated the statements only in two categories (common sense/non-common sense) and achieved an average accuracy of 51.0%. Even though we added an additional category in the evaluation of our method, we still achieved a higher positive average of 56.5%.

5. Conclusions and future directions

In this paper we presented a novel approach to automatically identifying common sense statements from unstructured texts and showed that it gave better results than previous methods.

Our ultimate goal is to create a semi-supervised agent for collecting and refining such statements. The agent will reside in the user’s browser. It will automatically identify statements as users browse and will engage the users in order to validate and/or refine the collected statements. With the help of user interaction we will be able to refine the category of vague statements (9.8%) as the user will be able to provide a much better understanding of the overall context in which the statement occurs. Thus, as far as the overall system is concerned, we can count both the positive average and vague average in the same category. Once we have perfected our approach, we plan to use the collected common sense to semantically annotate the World Wide Web.

However, in the course of this experiment we learned that doing the preprocessing of the text required a substantial amount of time and might not be feasible to be implemented as a real-time solution – an average of 21 sec/article for all the pre-processing steps described in 4.1. Thus, in addition to improving our method, we will be working on enhancing our preprocessing methods.

References

[Carlson 82] Carlson, G. N.: Generic Terms and Generic Statements, J. of Philosophical Logic, 11(2):145-181, 1982.

[Carlson 95] Carlson, G. N., F. J. Pelletier, editors: The Generic Book, University of Chicago Press, 1995.

[Carlson 99] Carlson, A., C. Cumby, J. Rosen, and D. Roth: The SNoW learning architecture, Technical Report UIUCDCS-R-99-2101, UIUC Computer Science Department, 1999.

[Fellbaum 98] Fellbaum, C.: Wordnet: An Electronic Lexical Database, Bradford Books, 1998.

[Lenat 90] Lenat, D.: Cyc:Towards Programs with common sense, Communications of the ACM, 33(8):30-49,1990.

[Liu 04] Liu, H., P. Singh: ConceptNet: A lexical database for English, BT Technology Journal, 4(22):211-226, 2004.

[Suh 06] Suh, S., H. Halpin, and E. Klein: Extracting Common Sense Knowledge from Wikipedia, Proceedings of the ISWC-06 Workshop on Web Content Mining with Human Language Technologies, 2006