社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－188 (16)
2008／11／27

Applying a contextual approach for collecting common sense statements to English and Bulgarian

Svetoslav DANKOV Rafal RZEPKA Kenji ARAKI
Graduate School of Information Science and Technology, Hokkaido University
E-mail: {dankov, rzepka, araki}@media.eng.hokudai.ac.jp

In this paper we present and build on a novel approach for automatically collecting common sense statements from the World Wide Web. As a backbone of our method we use generic rules and contextual clues to identify potential candidates. The generic rules consist of predetermined grammatical rules used to express common sense. The contextual clues consist of syntactic and semantic clues. The syntactic clues are represented by various syntactic structures frequently seen around common sense statements, while the semantic clues are represented by the various relationships between entities in the statement. To query for semantic relationships we are using WordNet. Two experiments were performed, evaluating the performance of our method, evaluating the viability of using semantic clues (WordNet) as well as the performance of our method when applied in another language (Bulgarian).

英語およびブルガリア語におけるコモンセンス表現収集のための文脈的アプローチの応用

スヴェトスラヴ ダンコヴ, ラファウ ジェプカ, 荒木 健治
北海道大学大学院情報科学研究科
E-mail: {dankov, rzepka, araki}@media.eng.hokudai.ac.jp

本稿では，Web 上のテキストデータからコモンセンス表現を収集する手法を提案する．我々は表現の候補を識別するために，一般性発見ルール及び文脈手掛かりを使用する手法を開発している．一般性発見ルールとは，常識的知識の出現しやすい文法上のパターンとしてあらかじめ定められたルールである．文脈手掛かりは文法的なものと意味論的なものとを含む．文法的手掛かりはコモンセンス表現の周辺に共通して存在する様々な文法構造として表現され，意味論的手掛かりはコモンセンス表現内の構成要素間の関係として表現される．意味論的関係を取り出すために WordNet を利用している．我々は，WordNet による意味論的手掛かりの有効性の評価，及び英語・ブルガリア語を用いてシステムの汎用性の性能評価を行った．

1 Introduction

One of the great challenges facing the artificial intelligence community is creating agents that can operate and adapt in a natural human environment. Naturally, we must be able to provide those agents with the information and the learning tools necessary for them to operate. If we want them to be able to make decisions about, relate to and have a simple understanding of the global environment in which they function, they need to be provided with basic knowledge about the world [1].

In humans, this knowledge is available to us in the form of reasoning shortcuts and factual information about each particular occasion we find ourselves in, and is generally known as common sense. We do not step out of 5th floor windows, since we know gravity would pull us towards the ground and we would probably die from the impact. We know that stepping in front of a moving vehicle would result in fatal injuries. We know that going to a birthday party entails buying a present and probably showing up early for the surprise party. We know these things because during our natural development, through both knowledge acquisition and reasoning, we naturally acquire the semantic values of objects around us and the relationships they form. 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.

Since most artificial intelligence systems of our age are very domain specific, and thus are able to operate within a very confined set of parameters, it

has been relatively easy to collect, represent and supply them with the necessary knowledge about their domains. Giving them common sense, however, explodes the size of information required to almost unimaginable degree. In the previous example, a robot would have to know the general axioms of physics (objects are subject to the force of gravity), its whereabouts (I am not on the first floor), general information about its surroundings (windows are not for exit, there is a door over there) and much, much more if we want it not to attempt a jump from that window.

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 first understand the general meaning of the text. In this paper we will show that there exist syntactic and semantic clues that can successfully help us identify common sense statements. We will also show that harvesting those semantic and syntactic clues with the help of generic rules can be successful in extracting common sense from natural language.

2 Background

2.1 A definition of common sense

Defining the term "common sense" is a difficult task. The term, unfortunately, coins a name for a phenomenon very hard to quantify or describe in detail. It has, however, captured the attention of AI researchers as early as 1959, when John McCarthy defined it in his seminal "Programs with Common Sense". He stated that "a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows" [2]. This definition proved instrumental in understanding and defining the requirements for artificial intelligence [1]. Marvin Minsky cunningly defined it as "uncommon sense" in his book "The Society of Mind". According to him, common sense is a collection of small bits of different types of knowledge, as opposed to a large amount of knowledge of only few varieties - in other words common knowledge as opposed to expert knowledge. For each domain of information we possess both a representation model and a body of skills to use on, or reason upon, that style of representation. Thus, representing common sense requires an impossible number of representation models and reasoning skills, as needed for the wide variety of fields common sense involves [3].

In sociolinguistic terms, "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 group. The information considered common sense in any culture includes many different variations and most times overlaps with the term cultural (or personal) beliefs.

From an engineering standpoint, however, those definitions are far too vague to merit a computational approach to common sense. What makes common sense difficult to work with is the fact that it does not simply represent information but also the result of a reasoning process about this information. Moreover, for humans, acquiring those representation models and reasoning processes is a natural and recursive process, during which we change and redefine those in order to best fit the situation. 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 Related research

With the realization of the importance of common sense to the field of artificial intelligence, considerable research has been done towards collecting and structuring this type of knowledge. The biggest research effort by far has been the Cyc project, which has already collected over a million common sense assertions in little over two decades. As the project became more of a commercial venture, a much smaller set of data is available free of charge. The work required, however, has been considerable. Common sense knowledge is manually input by experts in particular areas, who first give a complete ontological structure to the data, using a specially developed knowledge representation language called CycL, and then insert domain specific data based on their expertise [4]. A noted setback of the project is that the abstraction of upper ontologies is prone to loss of information, which in turn makes them highly speculative.

Another attempt to collect common sense data is the Open Mind Common Sense project. OMCS collects common sense statements from untrained volunteers over the Web in the form of natural language statements [5]. In the course of few years the project had already collected over 1.6 million statements. The major problem with OMCS, however, is that the data collected exhibited very low quality as there was no quality control. Current efforts in the new version of OMCS attempt to address those issues by providing stricter rules when users input data. 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 re-

searchers [6].

Very few researchers have attempted to develop automatic methods of collecting common sense statements with much success. The most important setback is the large amount of noise present when such methods are employed. The most notable attempt is focused on identifying the common sense orientation of noun phrases only as opposed to looking at the sentence as a whole, hence disregarding the context in which it appears. Later in our paper, we will show that our results surpass those attempts.

3 Our proposed method

In this section we will overview the main concepts in our method and give a detailed breakdown of it.

3.1 Genericity

As the core of our approach we employ the linguistic phenomenon termed "genericity" and the syntactic structures used to represent it. In the history of both philosophy of language and linguistics, there have been two distinct linguistic phenomena referred to as "genericity". The first is reference to a kind, as shown in example (1). In this instance the noun phrase "the lion" is a reference not to an individual instance of the object but a generalization over a particular group, namely the lion as species. Noun phrases like "lions" or "the lion" are called kind-referring noun phrases, or generic noun phrases, as opposed to object-referring noun phrases.

(1) "The lion is a major predator in Africa."

The second phenomenon is defined as propositions which do not express specific episodes or isolated facts, but instead report on a general property. These propositions generalize over particular episodes and facts, as opposed to generalizing over a kind as in the first case, and express characterizing properties. An example can be seen in (2). In this specific case the sentence is termed generic sentence as it expresses a generalization over a collection of episodes (Anna rode a bike yesterday, today and is likely to continue in the future). Clearly, this second notion of "genericity" is a semantic feature of the whole sentence, so it is very hard to harvest that information [7].

(2) "Anna rides a bike to work every morning."

For our purposes we use the kind-referring noun phrases as a starting point of our approach. We made the observation the statements starting with those are much more likely to contain common sense knowledge, thus we manually created rules that match such sentences and select them as possible candidates. In English, many definite singular count nouns, bare plural count nouns, and mass nouns can be considered kind-referring.

It is important to point out that one can view this kind of "generic" statements as a superset of common sense statements - predicates containing kind-referring noun phrases will not necessarily be common sensical (for example "Yesterday morning we shot the lions"). Therefore we need to look for other clues in order to refine our results [8]. In this experiment we also look at other contextual clues the details of which are explained below.

Table 1: Example set of syntactic keywords

Keyword Examples
Usual, usually
Common, commonly
Frequent, frequently
Typical, typically
Most, mostly
Every, all, some of, most
Always, never

3.2 Syntactic context

A further observation from the structure of generic statements was that there are particular syntactic features that can further strengthen the common sense value of those statements. As syntactic context we consider a specific set of adjectives and adverbs found in the noun group and verb group in the subject-verb-object relationship. Examples of this set are adjectives and adverbs like the ones seen in Table 1. Thus any sentence where either the noun group or the verb group has an adjective or adverb from that set will be selected as a common sense candidate.

3.3 Semantic context

In addition to looking at the syntactic context, we use the WordNet semantic database to look at the semantic context of the sentence. The WordNet database is a large lexical database of English, developed by George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms or synsets, each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations [9]. As WordNet has been a reliable source of week semantic relationships between words in English, we felt confident that it could serve as a great tool in establishing semantic relations in the 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. A graphical example of how we employ syntactic and semantic context can be seen in Figure 1.

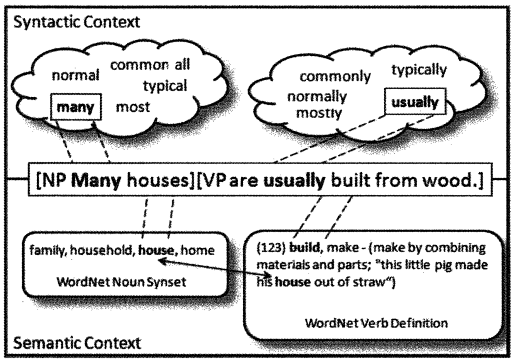

Figure 1: Illustration of syntactic and semantic context

3.4 Outline of 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 next steps we look at the syntactic and semantic contexts of the candidates. The method consists of the following eight steps and can be seen in Figure 2.

1. Input is parsed with a sentence splitter.
2. Words and sentences are tokenized.
3. The words are tagged with a part of speech tagger.
4. The sentence is chunked into noun chunks and verb chunk.
5. Generic rules are applied to each sentence to select common sense candidates.
6. Syntactic rules are applied to each sentence.
7. Each sentence's noun groups and verb groups are checked for semantic relations using WordNet.
8. Final candidates are extracted.

4 Evaluation Experiments

Our initial experiments were directed towards the feasibility of our method. Seeing that checking semantic context through WordNet was somewhat novel we decided to check if using it made a difference. We also checked how well our method could perform on another language - Bulgarian. Thus we devised three separate evaluation experiments: for our original method, testing in English; for our original method without using semantic context, testing in English; and for our original method without using semantic context, testing in Bulgarian (so far there is no freely available WordNet alternative for Bulgarian). As an agreement measure between users we used the Unweighted Cohen's Kappa measurement. The experimental setup and results for each of the experiments are described below.

4.1 English using semantic context

4.1.1 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, thus reducing noise to a minimum. During the random selection process, all articles were considered independent of their size or number of sentences they contained.

4.1.2 Evaluators

The extracted common sense statements were evaluated by two native speakers of English. Both are professionals, one involved in the IT sector, the other - a lawyer, both 29 years of age. The evaluators exhibited agreement of $k = 0.758$ during the evaluation, which shows a substantial agreement between them. They were asked to evaluate each statement candidate based on the following criteria.

1. If a statement is common sense - mark as "Yes".
2. If a statement is not common sense or noisy - mark as "No"
3. If the common sense value of the statement depends largely on the context in which it appears - mark as "Vague"

4.1.3 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 2, 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.

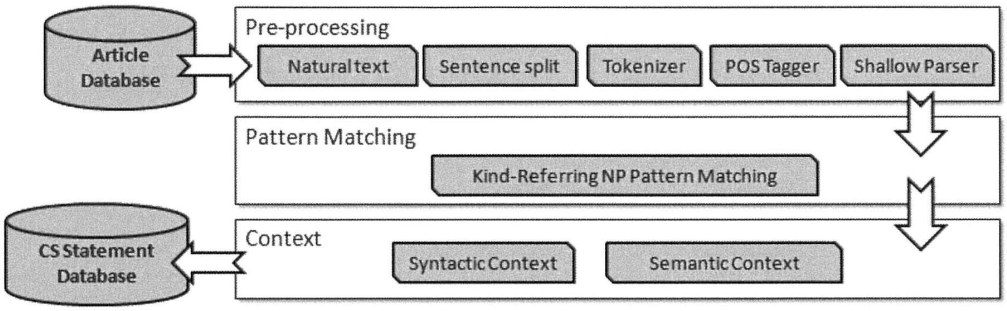

Figure 2: Outline of our method

Table 2: Results of evaluation experiment

		Eval. 2		
		Yes	No	Vague
Eval. 1	Yes	**635**	45	35
	No	50	**379**	11
	Vague	24	16	**110**

Table 3: Results of evaluation experiment

		Eval. 2		
		Yes	No	Vague
Eval. 1	Yes	**341**	18	25
	No	10	**363**	33
	Vague	31	12	**71**

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 [10] 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%.

4.2 English without semantic context

In this part of the experiment, the common sense candidates selected by the system were subjected to the same method as described in **3.4**, with the only difference being that the algorithm skipped step 7.

4.2.1 Corpus

For this part of the experiment, we used the same corpus and randomly selected articles as in **4.1**.

4.2.2 Evaluators

For this part of the experiment, we employed the same evaluators as in **4.1** with the same evaluation criteria. The evaluators exhibited agreement of k = 0.765 during the evaluation, which shows a substantial agreement between them.

4.2.3 Results

Out of the 16,475 articles, our algorithm found 892 common sense candidate statements in 408 separate articles. This represents 2.47% coverage on the original set of articles. The results are summarized in Table 3, 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 775.

Of the statements where both evaluators agreed, 44.0% were marked as common sense, 46.9% were marked as not common sense and 9.1% were marked as being too vague.

4.3 Bulgarian without semantic context

In order to run the experiment on Bulgarian texts, naturally we had to translate and adapt all the rules we had created for the method as described in **3**. We have taken a fairly na e approach, as there was not enough research done in the field of genericity in Bulgarian. The translation was, for the most part, a straight forward transition as most of the grammatical structures used to represent common sense in Bulgarian were very similar to those in English. There were a few notable exceptions that applied to Bulgarian, as described below.

1. Indefinite noun phrases carry generic meaning, as opposed to definite noun phrases in English

Table 4: Results of evaluation experiment

		Eval. 2		
		Yes	No	Vague
Eval. 1	Yes	**506**	96	61
	No	81	**539**	55
	Vague	17	39	**124**

2. Bare singular nouns can carry generic meaning as well

We also converted our syntactic keywords into Bulgarian and included ones specific to the language.

4.3.1 Corpus

For the purposes of our evaluation we are using the November 2008 snapshot of the Bulgarian XML Wikipedia article database and have selected all 62,000 articles. We have selected only the textual parts of those articles, discarding titles and any irrelevant information, thus reducing noise to a minimum.

4.3.2 Evaluators

For this part of the experiment, we used two native speakers of Bulgarian, a graduate student and an IT professional, 30 and 28 years of age respectively. We employed the same evaluation criteria as in **4.1** and **4.2**. The evaluators exhibited agreement of k = 0.623 during the evaluation, which shows a substantial agreement between them.

4.3.3 Results

Out of the 62,000 articles, our algorithm found 1,520 common sense candidate statements in 1,178 separate articles. This represents 1.9% coverage on the original set of articles. The results are summarized in Table 4, 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 1,169.

Of the statements where both evaluators agreed, 43.3% were marked as common sense, 46.1% were marked as not common sense and 10.6% were marked as being too vague.

5 Discussion

5.1 Full experiment

Our first experiment was designed to compare our method with the only other method for automatically collecting common sense statements we were aware of. As we mentioned in **4.1.3**, our method surpasses it even with one extra category of evaluation (the category for "Vague"). Compared to the 51.0% positive and 49.0% negative rate achieved in [10], we managed 56.5% positive and 33.7% negative, with 9.8% of those statements marked as vague. Even if all statements in the "vague" category were marked as negative we would still have a higher accuracy than the above method.

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 helping us harvest generic sentences. 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.

5.2 English without semantic context

In this experiment we discovered that semantic context, and WordNet in particular, proved valuable in the discovery of common sense statements. As we saw in **4.2.3**, only 44.0% (compared to 56.5% in **4.1.3**) were marked as common sense, 46.9% (compared to 33.7% in **4.1.3**) were marked as not common sense, and 9.1% we marked as vague. This confirms our initial suspicion that utilizing semantic information is vital in recognizing common sense.

5.3 Bulgarian without semantic context

In this experiment we attempted to naively translate our method and test it with another language. We are aware that common sense (in a cultural context) differs largely from country to country. However, our method in its current version does not rely on any cultural peculiarities. The results of this experiment are not very satisfactory: 43.3% of the statements selected by the system were marked as common sense, 46.1% were marked as not common sense, and 10.6% were marked as vague. The disappointing results could be attributed to the fact that the method was not specifically tailored for Bulgarian. It is interesting to note, however, that the results are comparable with those in **4.2.3**, which leads us to believe that if we had included a semantic context procedure we could have had much better results. Unfortunately, as of today there is no freely available WordNet for Bulgarian.

5.4 Conclusion

In this paper we discussed the importance of supplying common sense to artificial intelligence agents. We presented a novel approach to automatically identifying and extracting common sense statements from unstructured texts and showed that it gave better results than previous methods. We also showed the significance of using semantic information in recognizing common sense and the viability of using our method for another language.

References

[1] Lenat D. and E. Feigenbaum, "On the thresholds of knowledge", In Proceedings of International Joint Conference on Artificial Intelligence, William Kaufman, Cambridge, MA, 1987.

[2] McCarthy, J., "Programs with common sense", http://www-formal.stanford.edu/jmc/mcc95.html, 1959.

[3] Minsky, M., "The Society of Mind", Simon & Schuster, Inc., New York, 1985.

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

[5] Singh, P., "The public acquisition of common sense knowledge", In Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access, AAAI, Palo Alto, CA, 2002.

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

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

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

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

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