What is Hidden on the Web?
On Perspectives of the Statistically Based Estimation of the Web Content

Marcin Skowron and Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University
Japan
{ms,araki}@media.eng.hokudai.ac.jp

Abstract

This paper outlines the framework of the research on methods for automatic, language-independent categorization of the textual content of the Web. The focus of the presented work is the application of the lightweight frequency-based methods to distinguish between word collocations, utterances that convey commonsense knowledge, phrases that include less commonly known facts, and statements which are not reliable, or ill-formed. The foreseen utilization scope includes wide range of tasks such as automatic filtering and categorization of the knowledge concepts submitted by the volunteer contributors in the projects like Open Mind Commonsense or automatically retrieved from the WWW, and the categorization of user's utterances depending on their novelty-commonness for the purpose of the dialog and chat systems.

1 Scaling the Challenges to the Web

In eighties and nineties researchers successfully applied corpus based methods to address challenges, which at that time existed in the field of NLP. These attempts provided solutions or significantly improved the performance in various tasks including: Part-of-Speech Tagging, Parsing, Word Sense Disambiguation and Text Classification to name just a few of them. We believe that the rapid growth of the textual resources available on the Web along with the appearance of methods which effectively utilize this data, allow to address even wider spectrum of tasks, both from the Natural Language Processing and Artificial Intelligence fields. One of the biggest unresolved challenges in the field of AI is the task of providing commonsense knowledge to the computer systems. In the last decade there have been several project devoted to create large databases of commonsense knowledge concepts (Lenat 1995, Singh 2002). The examples of such knowledge concepts include: "apples are edible", "birds can fly" or " human beings need water to survive". The motivation of these projects is the belief that commonsense knowledge is vital to create truly intelligent systems. It is also foreseen that by providing a large number of knowledge concepts to the computer systems, they can perform reasoning in a way similar to human beings. For example, one of the envisaged application based on the commonsense knowledge database, the web search engine was able transform the novice Internet user's query "my cat is sick" to "veterinaries, Boston, MA" using chain of reasoning from the gathered knowledge facts (Liberman 2004). The number of knowledge facts necessary to create a database that could accommodate the amount of commonsense knowledge an average human being has, is usually estimated to include hundreds of millions axioms (Landauer 1986, Lenat 1990). After 20 years, this goal is far from being reached. We argue that the Web is a rich resource of commonsensical and general knowledge and that this resource is usable in the process of automatic creation of the knowledge databases. In the context of the knowl-

edge concepts retrieval the most important advantages of the Web include its scale and a wide coverage of various domains. At present, popular search engines index billions of web pages and this is just a fraction of the total number of pages available on the WWW. Assuming that only a small portion of the statements available on the Web can be treated as valid entries to a knowledge database, the Web still hosts a number of knowledge concepts that can be acquired in the amount unlikely to be obtained from any other text collection. The biggest challenge in the process of automatic acquisition of knowledge concepts from the Web is to obtain the high recall of valid knowledge concepts while ensuring the precision of their filtering and categorization.

2 Knowledge Concepts from Volunteer Contributors

Open Mind Commonsense (OMCS) is a widely known project aiming to collect knowledge concepts from the volunteer contributors using the Internet to facilitate the cooperation and knowledge submissions (Singh 2002). So far the OMCS project has gathered more than 700,000 items from more than 15,000 users. The OMCS database was evaluated using a sample[1] of the knowledge concepts, and the following conclusions were presented (Singh, Lim 2002): 75% of the items are largely true, 82% are largely objective, and 85% were judged as largely making sense. Table 1 shows the examples of knowledge concepts from the OMCS database.

3 Knowledge Concepts from the Web

Based on the observation that a substantive part of knowledge concepts submitted by the volunteer contributors in the OMCS project appears either as an exact match or in the paraphrased form on the WWW, the method for their automatic retrieval was proposed (Skowron 2005). Additionally, a large number of the statements available on the Web while constituting valid entries to a knowledge database, did not exist in the database created by the volunteer contributors. While the presented approach have the similar goal to this described in (Rzepka 2004), it significantly differs from the former work in the methods applied as well as the format of the generated database and the used language (English - Japanese). Our approach utilizes the clustering method to obtain a general, domain specific set of "web knowledge concepts candidates" and later based on the similarity measurement with the OMCS concept, performs the filtering and assigns the scores to the automatically retrieved "web knowledge concepts". The presented method was implemented in the "KnowY" system (Skowron 2005). Figure 1 presents the system flowchart[2].

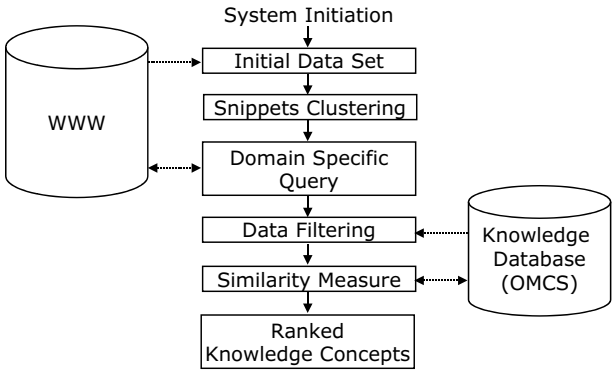

Figure 1: "KnowY" - The System Flowchart.

"KnowY" demonstrated that is was feasible to automatically obtain a large set of knowledge concepts that could be treated as the valid entries to a knowledge database. It also proofed that it was feasible to acquire the concepts even for the terms which were not covered in the OMCS database. On the other hand, similar to the knowledge concepts provided by the volunteer contributors, the content of the knowledge database generated by the "KnowY" was not error-free. Table 1 shows the examples of knowledge concepts from the KnowY database. As shown in the table, both the OMCS and KnowY databases consist of unsorted/uncategorized statements in the natural language form. These include valid commonsense knowledge concepts, statements describing personal opinions or specific situations, and phrases which are unreliable or ill-formed. Although, some of the applications that used the OMCS database demonstrated that it was feasible to reason from this noisy set (Eagle 2003,

[1]The sample was obtained after eliminating the items marked by the human judges as garbage. Such items accounted for 12.3% of the randomly generated sample (Singh, Lim 2002).

[2]For the detailed method and implementation description please refer to (Skowron 2005).

Table 1: Examples of knowledge concepts from the OMCS and KnowY databases.

OMCS Concept	OMCS Concept	KnowY Concept
One plus one is two	Half a glass of water	Earth has one moon
A bottle contains liquid	I build a house	Apples have 5 seeds
An admission ticket	Bus stops are usually dirty	MacIntosh type apple
The sun feels warm	An Iexporter can ship goods	Water is remarkable
Apples are green	Angels do not exist	My wife's favorite early apple
The Sun feels warm	Jason got a stomach ache	We live on the planet Earth

Stocky 2004), for various other ones the additional refining, verification and categorization of OMCS and KnowY databases is highly desirable.

4 Commonness Estimation

The ability to automatically discover the word collocations and proverbs from a large set of documents is widely known (Manning and Shutze 1999). In this process a relatively simple, frequency based analysis lead to automatic acquisition of large number of collocations such as "kick the bucket", or expressions like "on the other hand". Such phrases can be characterized by the strength of links between the words from which they are constructed. The presented approach aims to extend the application scope of the frequency based methods to roughly categorize a given phrase depending on its character to the one of the following classes: word collocation, valid or invalid knowledge concept[3], and unreliable or ill-formed phrases. In the presented approach we used formula **1** (PS - PhraseScore), to assign a score for a given phrase based on the Web statistic of n-grams composing this phrase.

$$PS = log\frac{\sum_{n=1}^{N} F(W_n)}{\sum_{n=1}^{N} F(W_n + W_{n+1})} * FS * W_c \quad (1)$$

, where $F(W_n)$ - frequency of a word, $F(W_n + W_{n+1})$ - frequency of a bi-gram, $F(W_1 + ... + W_N)$ - frequency of a given phrase, FS - $log(F(W_1 + ... + W_N)$, W_c number of words in a phrase. For the n-grams for which occurrence statistic was not discovered, the frequency 1 was assigned[4].

[3]The concepts classification is based in the evaluation criteria as described in (Singh, Lim 2002).

[4]Frequency counts were obtained from the Google search engine: http://search.google.com.

5 Experiments

For the preliminary experiments, 100 statements not longer than 5 words were randomly selected from the OMCS. The examples of the ranked concepts along with their PhraseScore are presented in Table 2. The PhraseScore higher than 0 was assigned to 52 concepts found in the used sample of the OMCS database. Although, no specific thresholds were defined, we found the intuitive gradation between the concepts, depending on their PhraseScore. The examples 1-5 provided mostly word collocations and commonsense knowledge concepts. The examples on positions 6-52 presented the combination of the commonsense and general knowledge facts. The remaining statements (53-100) for which the PhraseScore equaled 0, contained the description of specific situations or individual opinions as well as the less frequently known facts. Other statements from this set were either misspelled or ill-formed grammatically.

6 Conclusions and Future Work

This paper presented the outline of the research on the methods for the automatic categorization of phrases found on the Web. The proof of the concepts was performed on the set of concepts from the OMCS database, demonstrating that it is feasible to automatically validate and roughly categorize the submitted statements depending on the character of a given phrase. This information can be used to distinguish between word collocations, utterances that convey commonsense knowledge, phrases that include less commonly known facts, and statements which are not fully reliable, or ill-formed grammatically. The refining of the method and the precise settings of the thresholds necessary to apply the method

Table 2: Examples of OMCS concepts with the corresponding PhraseScore.

No.	OMCS Concept	PhraseScore	No.	OMCS Concept	PhraseScore
1	An admission ticket	188.66	40	Hanukkah is a Jewish Celebration	33.14
2	Half a glass of water	176.19	41	Communism is based on Socialism	30.34
3	It takes all kinds	159.05	42	Dogs like to play fetch	29.97
4	One plus one is two	155.10	43	Lettuce is a vegetable	29.38
	...			...	
12	The Sun feels warm	99.42	52	Food can be very pleasurable	13.28
13	Some flowers are yellow	95.42	53	Quimper is in Britanny	0
14	Leaves are usually green	82.89	54	Sometimes drinking causes is hydration	0
15	Hats were once very popular	74.37	55	Jason got a stomach ache	0

in the larger scale are envisaged as our future work. There is a wide range of applications that could potentially benefit from fully developed commonness estimation method, including story generation, machine translation (text post-processing), dialog and chat systems (recognizing and categorizing the users utterances; commonness-novelty estimation), automatic knowledge acquisition and verification (including the correction of misspelled and grammatically ill-formed knowledge concepts).

References

Eagle N., Singh P. and Pentland A. Common sense conversations: understanding casual conversation using a commonsense database. *Proceedings of the Artificial Intelligence, Information Access, ad Mobile Computing Workshop(IJCAI 2003)*, 2003.

Landauer T. How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science, 10(4)*, pages 477-493, 1986.

Lenat D. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM, 38(11)*, pages 33-38, 1995.

Lenat D., Guha K., Pittman K., Pratt D. and Shepherd M. CYC: towards programs with common sense. *Communications of the ACM, 33(8)*, pages 30-49, 1990.

Lieberman H., Liu H., Singh P., and Barry B. Beating common sense into interactive applications. *AI Magazine*, 25(4),pages 63-76, 2004.

Manning C. and Shutze H. Foundations of Statistical Natural Language Processing. Massachusetts Institute of Technology, 1999.

Rzepka R., Itoh T. and Araki K. Rethinking Plans and Scripts Realization in the Age of Web-mining. *IPSJ SIG Technical Report 2004-NL-162*, pages 11-18, 2004.

Singh P. The public acquisition of commonsense knowledge. *In Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, 2002.

Singh P., Lim G., Lin T., Mueller E., Perkins T., Tompkins M. and Zhu W. Open Mind Common Sense: Knowledge Acquisition from the General Public. *Proceedings of the Fifth International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, 2002.

Skowron M., Araki K. Voluntary Contributions of Unaware Internet Users. On Automatic Knowledge Retrieval from the WWW. AAAI 2005 Spring Symposium, Knowledge Collection from Volunteer Contributors. 2005.

Stocky T., Faaborg A. and Lieberman H. A Commonsense Approach to Predictive text Entry. *Conference on human Factors in Computing Systems (HI 04)*, 2004.