Extracting Comparative Commonsense from the Web

Yanan Cao, Cungen Cao, Liangjun Zang, Shi Wang, Dongsheng Wang

Extracting Comparative Commonsense from the Web

Yanan Cao[1,2], Cungen Cao[1], Liangjun Zang[1,2], Shi Wang[1], Dongsheng Wang[1,2]

[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
No.6 Kexueyuan South Road Zhongguancun,
Beijing 100190, China
[2] Graduate University of Chinese Academy of Sciences
No. 19 Yu Quan Road, Shi Jing Shan Distinct,
Beijing 100049, China

Abstract. Commonsense acquisition is one of the most important and challenging topics in Artificial Intelligence. Comparative commonsense, such as "In general, a man is stronger than a woman", denotes that one entity has a property or quality greater or less in extent than that of another. This paper presents an automatic method for acquiring comparative commonsense from the World Wide Web. We firstly extract potential comparative statements from related texts based on multiple lexico-syntactic patterns. Then, we assess the candidates using Web-scale statistical features. To evaluate this approach, we use three measures: coverage of the web corpora, precision and recall which achieved 79.2%, 76.4% and 83.3%, respectively in our experiments. And the experimental results show that this approach profits significantly when the semantic similarity relationships are involved in the commonsense assessment.

1 Introduction

Commonsense knowledge plays an important role in various areas such as natural language understanding (NLU), information retrieval (IR), question answering (QA), etc. For decades, there has been a thirst in artificial intelligence research community for large-scale commonsense knowledge bases. Since the hand-coded knowledge base suffers from semantic gaps and noises, and the human effort should still be involved in the maintenance [1], much recent work focuses on automatic strategies to acquire commonsense knowledge.

As information people use every day, commonsense knowledge encompass many aspects of life. It relates to an individual object, phenomena and activity, or relations between different concepts and a sequence of events. This paper concentrates on a commonsensical concept-level relation, which is defined as follows.

Comparative Relation We say that concept c_1 and c_2 are comparative if and only if the natural language description "In general, c_1 is relatively * than c_2" is reasonable and acceptable, where "*" is an adjective expressing the comparative degree. This statement denotes that the concept c_1 has a property

or quality greater or less in extent than that of c_2. We use the triple (c_1, c_2, p) to express it formally. For a specific adjective in this statement, we call it a comparative property of the given pair of concepts. For example, "In general, a man is relatively stronger than a woman" describes a comparative relation between the concepts man and woman, of which *stronger* is the comparative property. The formalization of this statement is (*man*, *woman*, *stronger*).

In this paper, we aim to acquire comparative commonsense about given pairs of concepts from the Web, which is a vast source of information. Firstly, we submit instantiated lexico-syntactic patterns as query terms to the search engine, such as "men are * than women", and extract potential comparative statements from the retrieved texts. Then, we assess the candidates based on multiple Web-scale statistical features. The semantic similarity relationships, including synonym and antonym, are involved in the computing of these statistics. This improves the effectiveness of our approach. In the experiments, we selected twenty pairs of concepts as the input data and the result set of comparative statements had a precision of 76.4% and a recall of 83.3%. Besides, we show the high coverage of the comparative information in a test web corpus, which achieved 79.2%.

The remainder of this paper is organized as follows. In Section 2, we review some related work on commonsense knowledge acquisition from texts. Section 3 and Section 4 describe the two-stage method for extracting and assessing comparative commonsense in detail. In Section 5, we evaluate our approaches and discuss the experimental results. Finally, we conclude in Section 6.

2 Related Work

To acquire commonsense knowledge, a number of efforts tap on textual data sources. The pioneering work mentioned in [2] recognized text as a source of both explicit and implicit commonsense knowledge. Capitalizing on interpretive rules, they obtained a broad range of general propositions from noun phrases and clauses. And they tried to derive stronger generalizations based on the statistical distribution of the obtained claims. However, to design abundant rules is indeed time-cost and the extraction process was not implemented automatically.

[3] extracted commonsensical inference rules from coordinated verb phrases in Japanese text. This work was based on the assumption that if two events shared a common participant, then the two events would probably be a probably-follow relation. They used an unsupervised approach to select the inference rules based on lexical co-occurrence information. The results were evaluated by multi-subjects, and just 45% of the selected rules were accepted by 80% of the human subjects. The low precision showed that co-occurrence is important but not adequate to assess the specific causal relation between two events, because there are other event-level relations such as *IsPreferableTo* in [4].

[4] used the Web data to clean up and augment existent commonsense knowledge bases. For specific binary event-level relations such as causality, they used complete or partial event descriptions and lexico-syntactic patterns to mine potential instances of the relations, and assessed these candidates based on Web-

scale statistics. For higher-level predicates, the relational definitions were involved in the assessment, which extended the work in [3]. The advantage of this work is that some implicit instances of relations can be assessed at the level of the Web rather than found in one particular sentence.

In our recent work [5], we acquired commonsense knowledge about properties of concepts by analyzing how adjectives are used with nouns in everyday language. We firstly mined a large scale corpus for potential concept-property pairs using lexico-syntactic patterns and then filtered noises based on heuristic rules and statistical approaches. For each concept, we automatically selected the commonsensical properties and evaluated their applicability. The precision of the result-set achieved 67.7%. In the following, we continue our research on the relations between concepts and properties, from the view of comparing.

3 Extracting Candidate Comparative Statements

This section describes the first phase of acquiring comparative commonsense: we extract candidate commonsense from the Web using linguistically-motivated patterns [6]. In the text, comparative relations are expressed in various forms. Table1 lists several patterns followed by corresponding instances, in which *NP* denotes *Noun Phrase*.

Table 1. Examples of patterns designed to extract comparative commonsense

Pattern	Instance
NP_1 be * than NP_2	men are stronger than women
NP_1 be more * than NP_2	men are more powerful than women
NP_2 be not * than NP_1	women are not stronger than men
NP_2 be less * than NP_1	women are less powerful than men
to compare NP_1 and NP_2, NP_1 be *	to compare men and women, men are stronger
compared with NP_2, NP_1 be *	compared with women, men are stronger

When we perform the extraction process, these patterns are automatically instantiated with given pairs of concepts to generate query terms. For example, to acquire comparative statements about *man* and *woman*, we instantiate the first pattern as "men are * than women" and the last one as "compared with women, men are *". Issuing the query terms, we take advantage of the search engine to retrieve relevant texts from the Web and obtain the hit count of each query string. And subsequently, we extract a comparative property from each snippet matched with the "*" wildcard, which constitutes a candidate triple with its corresponding concept-pair.

There are two explicit constrains on the extraction. The first is that the matching snippet should be an adjective phrase, and we just extract the head

word no matter it has a modifier or a complement. For example, from the text "men are much stronger than women", we extracted the comparative property "stronger", while "much" is discarded. We implement this constraint by analyzing the combination of POS in adjective phrases. Second, we extracted the comparative statements in a context-independent manner. That means we don't extract the specific context even if a statement is not acceptable anymore without it. For example, in the sentence "From Shanghai to Lhasa, a train is faster than an airplane", the extracted statement (*train*, *airplane*, *faster*) is dependent on the context "from Shanghai to Lhasa", or else it violates the general knowledge. To avoid obvious errors induced by this strategy, we restrict the matching snippet to be a single sentence or to have a common modifier such as "generally speaking", "as we all know", etc.

4 Assessing Candidate Comparative Statements

During the extraction process, we acquire all potential comparative statements. However, the Web contains massive unreliable or biased information [7], and some knowledge we obtained is inconsistent with the commonsensible facts. So, we verify each candidate statement based on multiple Web-scale statistics.

4.1 Statistical Features for commonsense assessment

Feature 1 Occurrence Frequency. Because a frequently mentioned piece of information is typically more likely to be correct than an item mentioned once or twice, high frequency is an important feature for truth. Given a candidate statement (c_1, c_2, p), we could easily obtain its occurrence frequency $Freq(c_1, c_2, p)$ in the web corpora using hit counts of instantiated queries, which are returned by the search engine:

$$Freq(c_1, c_2, p) = \sum\nolimits_{pt \in PT} hits(pt(c_1, c_2, p)), \tag{1}$$

where PT is the set of predefined patterns (referred in Table 1) and $hits(pt(c_1, c_2, p))$ represents the hit count for a query which is an instance of the pattern pt.

Feature 2 Confidence. Although frequency is an important feature, it's not a guarantee of truth. Given a pair of concepts c_1 and c_2, we use confidence to indicate the probability of a property p to be their comparative property. More specifically:

$$Conf(c_1, c_2, p) = Freq(c_1, c_2, p) / \sum Freq(c_1, c_2, p) \tag{2}$$

In this formula, the denominator is the number of potential comparative relations which c_1 and c_2 participant in, and "*" is any potential comparative property of c_1 and c_2.

Feature 3 the Number of the Matched Patterns. The last feature is the number of different expression forms of the same comparative statement. It's motivated by the assumption that if a potential statement is extracted by multi-patterns, it seems more reliable.

4.2 Semantic Similarity Relationships in the Candidate Set

We note that, the candidate statements are not independent, and there are semantic similarity relationships among them. For example, (*woman*, *man*, *weaker*) is consistent with the statement (*man*, *woman*, *stronger*), while (*man*, *woman*, *weaker*) is opposed to it. We make use of these semantic similarity relationships in the computing of the statistics. To see this, we first begin by introducing several fundamental notations.

Given a potential comparative statement (c_1, c_2, p), we use $SYN_p = \{x|x \text{ is a synonym of } p\} \cup \{p\}$ to denote the set of all synonyms of the property p, and another set $ANT_p = \{y|y \text{ is a synonym of } p\}$, which contains all antonyms of p. Based on these two sets, we divide other statements into the following categories according to the relationship between it and (c_1, c_2, p).

- A **supporting example**, which has the same meaning of the statement (c_1, c_2, p). For instance, (*man*, *woman*, *more powerful*) and (*woman*, *man*, *weaker*) are both supporting examples of (*man*, *woman*, *stronger*). It belongs to the set

$$SupSet(c_1, c_2, p) = \{(c_1, c_2, p')|p' \in SYN_p\} \cup \{(c_2, c_1, p''|p'' \in ANT_p\}$$

- A **counterexample**, which is opposed to the statement (c_1, c_2, p) in semantic. Using above instance (*man*, *woman*, *stronger*), (*man*, *woman*, *weaker*) and (*woman*, *man*, *stronger*) are both its counterexamples. The set of the counterexamples is

$$CntSet(c_1, c_2, p) = \{(c_1, c_2, p')|p' \in ANT_p\} \cup \{(c_2, c_1, p''|p'' \in SYN_p\}$$

- An **irrelative example** is the statement that is neither a supporting example nor a counterexample. For example, (*man*, *woman*, *braver*) and (*man*, *woman*, *stronger*) are irrelative to each other.

Then, we use $Support(c_1, c_2, p)$ to denote the frequency of supporting examples of the given triple (c_1, c_2, p). More specifically:

$$Support\,(c_1, c_2, p) = \sum\nolimits_{(c_1, c_2, p') \in SupSet(c_1, c_2, p)} Freq\,(c_1, c_2, p') \tag{3}$$

And we use $Counter(c_1, c_2, p)$ to denote the frequency of counterexamples of the given triple (c_1, c_2, p). It is computed as follows:

$$Counter\,(c_1, c_2, p) = \sum\nolimits_{(c_1, c_2, p') \in CntSet(c_1, c_2, p)} Freq\,(c_1, c_2, p') \tag{4}$$

Intuitively, if supporting examples of the statement (c_1, c_2, p) are much more than its counterexamples, this statement is more likely to be true. So, we use $ExtFreq(c_1, c_2, p)$ instead of $Freq(c_1, c_2, p)$ to assess potential comparative statements:

$$ExtFreq\,(c_1, c_2, p) = Support(c_1, c_2, p) - Counter(c_1, c_2, p) \tag{5}$$

And confidence is accordingly computed as follows:

$$ExtConf(c_1, c_2, p) = ExtFreq(c_1, c_2, p) / \sum Freq(c_1, c_2, *) \tag{6}$$

Although the features frequency and confidence are not novel, we combine Web-scale statistics and semantic relationships, and this improves the performance of our approach.

5 Experiments

In this section, we report some experimental results on comparative commonsense acquisition from Chinese web corpora. We computed the estimated coverage to prove the Web a proper resource for comparative commonsense acquisition. And we show that using similarity relationships in the assessment improves the precision and recall of our approach.

5.1 Experimental Settings

We selected twenty pairs of our familiar nouns as the test concept-set, which were averagely divided into four classes: human, plant, animal (except for human) and artificiality. These concepts were all basic level categories [8] (such as "plane"), which are neither too abstract (such as "vehicle") nor too concrete (such as "jet plane"). This is because the concepts on basic level are more likely to be proper for commonsense acquisition [9]. Intuitively, not arbitrary pairs of concepts are comparable. They always have properties in common, but one's property is greater or less in extent than that of the other. To satisfy this assumption, we picked out a concept-pair from the same class, such as "man" and "woman", "watermelon" and "apple", rather than "man" and "apple".

Next, we issue each concept-pair to the Google Search Engine and downloaded relevant texts the search engine returned from the Web. We call it the *Test Corpus* in the following experiments.

5.2 Results: Coverage of the Web Resource

Because an average person is assumed to possess commonsense knowledge, it is not communicated most of the time. That means commonsense knowledge is usually used implicit in the texts. However, it seems infeasible to extract knowledge implicit in texts just using the pattern-matching method. To investigate

what amounts of comparative knowledge we might extract from explicit expressions on the Web, we propose the coverage of comparative information on the web:

$$Coverage = \frac{the\ number\ of\ comparative\ statements\ explicit\ on\ the\ Web}{the\ number\ of\ comparative\ statements}$$

In fact, we don't know the accurate quantity of commonsense possessed by people or that of information distributed on the whole web. So we estimated these values based on several human subjects and the *Test Corpus*. More specifically:

$$EstCoverage = \frac{the\ number\ of\ comparative\ statements\ in\ the\ Test\ Corpus}{the\ number\ of\ comparative\ statements\ from\ human\ subjects}$$

We asked five human subjects to contribute their comparative commonsense on the test concept-set. The subjects' knowledge was gathered in an iterative way. Firstly, they were limited to derive this knowledge just from their brains. After the subjects submitted their outcome for the first time, we presented them the *Test Corpus*. In the second phase, the subjects manually extracted comparative commonsense from the corpus, which may provide them new clues. Meanwhile, the subjects were allowed to modify their primary submission.

Table 2. Quantity of comparative commonsense obtained from 5 human subjects and that in the test web corpus

Noun Class	Human Subjects		Test Corpus	Coverage
	primary	modified		
Human	99	128	102	79.7%
Plant	27	33	30	90.9%
Animal	57	57	31	54.4%
artificiality	50	61	58	95.1%
total	233	279	221	79.2%

Table 2 shows the quantity of the twice submissions of the subjects and knowledge automatically mined from the web corpus, respectively. We note that, the subjects supplemented 46 records according to the web resource, which were not referred in the first submission. And the average coverage achieved 79.2%, which demonstrated the Web a good resource containing adequate explicit comparative commonsense.

5.3 Results: Effectiveness of Our Approach

In this experiment, we used the feature *Freq* and *Conf* to assess potential statements as the baseline method, and used *ExtFreq* and *ExtConf* as the extended

method. Precision and recall are used to evaluate the performance of these approaches.

Before our extraction, we pre-processed the related texts in the test corpus including HTML label analyzing, word segmentation and POS tagging [10]. We extracted 429 potential comparative statements using the heuristic rules (referred in Section 3.1). Then, we expressed these acquired triples in the natural language "In general, c_1 is relatively * than c_2", and asked the five subjects to judge whether each description was acceptable. If a potential statement was accepted by 60% of the human subjects, we consider it a piece of comparative commonsense. Finally, we assessed each potential statement using the statistical features in both baseline method and extended method, and the results were compared with that from the human judgments.

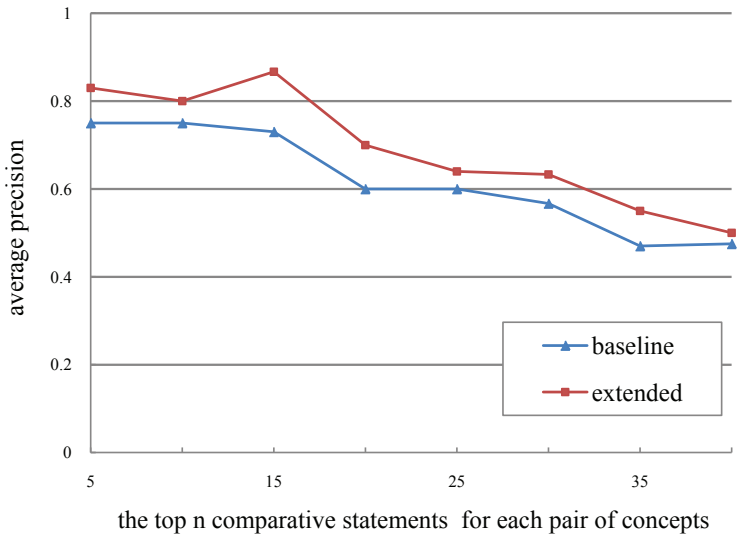

Fig. 1. Comparison of two methods

Fig.1. shows the performance of both the baseline method and extended method. Given the minimum threshold of confidence and the number of matched patterns which are 0.057 and 3 respectively, we ranked the comparative statements according to their frequencies. This graph plots the average precision and the top n acceptable statements ranked for a pair of concepts. It's obvious that the extended method outperformed the baseline scores. The average precision and recall achieved 76.4% and 83.3%, respectively.

6 Conclusion and Future Work

Comparative commonsense describes the concept-level relations from the view of comparing, which denotes that one entity has a property or quality greater or less in extent than that of another. In this paper, we propose an automatic method for acquiring comparative commonsense from the World Wide Web. We use multiple lexico-syntactic patterns to extract potential comparative statement from related texts retrieved by the search engine. And then, we assess the candidates by combining Web-scale statistics and their semantic similarity relationships including synonym and antonym.

In the experiments, we evaluated the coverage of the explicit comparative information in the web corpus, which demonstrated the Web a good resource for our extraction. The experimental results also showed that the use of semantic similarity relationships in assessment significantly improves precision and recall of our approach when the statements have close dependency.

This work is based on given comparative pairs of concepts. In the future work, we will research on the characteristics of the comparative concepts, and automatically select them as the input data.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant No. 60773059.

References

1. Singh P.: The Public Acquisition of Commonsense Knowledge. In Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access. 2002
2. Schubert L.: Can We Derive General World Knowledge from Texts. In Proceedings of the Second International Conference on Human Language Technology. 2002
3. Torisawa K.: An Unsupervised Learning Method for Commonsensical Inference Rules on Events. In Proceedings of the Second CoLogNet-EIsNET Symposium. 2003
4. Popescu A.M.: Information Extraction from Unstructured Web Text. Ph.D. thesis, University of Washington. 2007
5. Cao Y.N., Cao C.G, Zang L.J., Zhu Y., Wang S., Wang D.S.: Acquiring Commonsense Knowledge about Properties of Concepts from Text. In Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery. **4** (2008) 155-159
6. Hearst M.A.: Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International Conference on Computational Linguistics. (1992) 539-545
7. Kosala R., Blockeel H.: Web mining research: a survey. SIGKDD Explorations. **2** (2000) 1-15
8. Zhang M., Cognitive Linguistics and Chinese Noun Phrases, China Social Science Press, Beijing. 1998

9. Zhu Y., Zang L.J., Cao Y.N., Wang D.S., Cao C.G.: A Manual Experiment on Commonsense Knowledge Acquisition from Web Corpora. In Proceedings of the 7th International Conference on Machine Learning and Cybernetics, Kunming. 2008
10. Zhang H.P., Yu H.K., Xiong D.Y., and Liu Q.: HMM-based Chinese Lexical Analyzer ICTCLAS. In Proceedings of the second SIGHAN Workshop affiliated with 41st Annual Meeting of the Association for Computational Linguistics, Sapporo Japan. 2003