Research Article

Paraphrasing Chinese Idioms: Paraphrase Acquisition, Rewording and Scoring

Jia Jun, Dong[1], Tien-Ping, Tan[2*]

[1,2]School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
[2]tienping@usm.my

Article History: Received: 11 October 2020; Accepted: 27 December 2020; Published online: 8February 2021

Abstract: Paraphrasing is a process to restate the meaning of a text or a passage using different words in the same language to give a clearer understanding of the original sentence to the readers. Paraphrasing is important in many natural language processing tasks such as plagiarism detection, information retrieval, and machine translation. In this article, we describe our work in paraphrasing Chinese idioms by using the definitions from dictionaries. The definitions of the idioms will be reworded and then scored to find the best paraphrase candidates to be used for the given context. With the proposed approach to paraphrase Chinse idioms in sentences, the BLEU was 75.69%, compared to the baseline approach that was 66.34%.

Keywords:paraphrasing, idioms, multi-word expressions

1. Introduction

Idiom is an expression which is the crystallization of human wisdom. It consists of a few words, but they often tell a story, and they can be admonition of wisdom. Many languages have idioms. Most of the idioms in Chinese consists of four characters and they are known as chengyu (成语). For example: 三顾茅庐, which means “visiting the thatched hut three times” literally. It was based on an event during the Three Kingdoms period in China, where Liu Bei, a warlord visited and persuaded Zhuge Liang, a renowned military strategist, three times in order to recruit him. One of the meaning of this idiom is “wholehearted”. There are no official numbers of chengyu in Chinese. The chengyu dictionary 汉语成语 (2018) documented 8 thousand idioms, while some chengyu dictionaries have more than 20 thousand entries, such as 新华成语大词典 (2017) and 20000 条成语大词典(全新版) (2016).

Paraphrasing is a process to restate the meaning of a text in a different form, often, the purpose is to give readers a clearer understanding. After paraphrasing a text, the text will contain almost the same information. Paraphrasing can be done by changing the sentence structure, for example changing from active voice to passive voice, or/and substituting some text with a different text that has the same meaning. For example, one of the paraphrases for the sentence (i) is the sentence (ii).

(i) he is a lecturer.
(ii) he teaches at a university.

The phrase “... is a lecturer” can be paraphrased as “... teaches at a university”. If there is a new sentence, e.g. “Dr. Koehn is a lecturer.”, it can restate as “Dr. Koehn teaches at a university.”.

Idioms are interesting expressions because often they cannot be understood literally. Consider the idiom “stick to your guns”, which cannot be understand just from the words. In cases where someone does not understand an idiom in a sentence, the sentence has to be put in another way to let the reader understand it. In this article, we propose an approach to paraphrase Chinese idioms in sentences by using definitions found in idiom dictionaries as paraphrase candidates. A scoring approach is then used to evaluate the paraphrases and select the one that best match the given context. We also combined the proposed paraphrasing with an MT and evaluate the translation produced.

2. Acquiring Paraphrases

Paraphrase corpus is a resource used to create paraphrase model for recognition and generation of paraphrases. Many researchers have attempted various methods to build a paraphrase corpus, and this section describes some of the works in this area.

2.1 Lexical corpus

Synonym is a word that has the same or similar meaning as another word in the language. Thus, one way to paraphrase a text is to substitute words in the text with synonyms. Synonyms can be extracted from dictionaries. Another source of synonyms is WordNet (Fellbaum, 1998). Words are grouped into sets of near-synonyms called synsets, and hypernyms are arranged in hierarchy in WordNet. The lexical database provides six measures of similarity, and three measures of relatedness between a pair of synsets (Pedersen et al., 2004). There are also many works that extend the functionality of WordNet to measure the similarity between sentences (Mihalcea et al., 2016).

Pershina et al. (2015) attempted to identify the paraphrase of an idiom using a similar idiom. They compared the similarity (in meaning) of two English idioms by using the definitions of the idiom. The lexical similarity and the semantic similarity were calculated from their definitions. The lexical similarity is calculated based on the cosine similarity between word embedding vectors. The semantic similarity captures the difference in the overall semantic meaning of the definitions by combining the word embedding vectors and then calculate their difference using the cosine similarity. The approach may also be applied on non-idiom phrases. However, paraphrasing an idiom with a non-idiom phrase will require extra steps to make sure the resulting sentence valid.

2.2 Parallel translations

Some popular texts are translated by different authors to the same target language. These translated texts serve as a good resource for acquiring paraphrases. A study was done on novels (Barzilay and McKeown, 2001). The sentences from different texts are first aligned. The paraphrases are then extracted based on the idea that the phrases in aligned sentences that appear in similar context are paraphrases. To determine the context that contains paraphrases, contexts with identical words in aligned sentences are extracted and filtered according to their predictive power. Paraphrases are extracted from these contexts. The approach then uses the extracted paraphrases to learn general syntactic patterns to learn new context rules. The advantage of this method is the high quality of the obtained paraphrases. The disadvantage is that resource acquisition is difficult, and the parallel sentences that meet the criteria are very rare, and only some phrases can be obtained.

2.3 Comparable text

Comparable texts are texts with similar content. Comparable texts can be in different languages, but for extracting paraphrases, monolingual comparable texts are required. Furthermore, there is no agreement on the minimum amount of similarity that must exist between comparable text. Paraphrase can be viewed as a special case of comparable text, where the text has the same or nearly the same meaning. Thus, extracting paraphrases from comparable text involves identifying words, phrases or sentences that have the same meaning. There are some studies to extract paraphrases from comparable texts such as news articles, since many different articles that report on the same eventis available. The Microsoft Research Parallel Corpus (Dolan and Brockett. 2005) was extracted from comparable texts. Some heuristics such as lexical similarity from minimum edit distance, sentence position in a document, and sentence length were used to first locate possible paraphrase pairs from similar news articles. An SVM binary classifier trained using a manually annotated paraphrase corpus was then used to separate the paraphrase/non- paraphrase pairs.

2.4 MT Approaches

There are also attempts to use bilingual parallel corpora to induce paraphrases by using phrase-based statistical machine translation (SMT) (Koehn et al., 2007). The translation model of a phrase-based SMT consists of a phrase translation table. To find the paraphrase for a phrase (e1), other phrases (e2) with the same translation/phrase (f) from the phrase translation table have to be lookup. In other word, the approach uses the transitive relation (if A is related to B, B is related to C, then A is related to C) to find paraphrase. The idea is subsequently extended to syntactic paraphrases that lead to the creation of a large paraphrase corpus that consists of more than a billion pairs of paraphrases in 23 languages.

3. Paraphrase Generation

A paraphrase corpus is important for modeling paraphrases. A paraphrase corpus contains examples of phrases and their respective paraphrases. The paraphrases collected can be used to generate paraphrase given a text. Paraphrase generation is not simply about substituting a text with another text because the resulting text may be wrong grammatically, particularly when the text one is substituting is a phrase instead of a sentence. The text may need to be edited after the substitution. Besides, there may be more than one way to paraphrase a text. Selecting the right one for the context is also a challenge.

3.1 MT Approaches

Translation and paraphrasing are very similar. Translation renders the idea of a text in a different language, while paraphrasing renders the same idea in the same language using different words. Thus, theoretically a machine translation (MT) system can also be used for paraphrasing given paraphrase examples. Among the state-of-the-art machine translation systems are statistical machine translation (SMT) and neural machine translation (NMT). Both are data-driven approach, where the translation model is trained using a parallel text corpus. Nevertheless, the paraphrase corpus is either not exist or it is very small. For instant, the English-Chinese UN parallel corpus consists of more than 15 million parallel sentences (Michal et al., 2016), compare to the large-scale paraphrase corpus, Microsoft Research Paraphrase Corpus consists of only more than 5 thousand sentences.

3.2 Pivot Language Approaches

Another way to paraphrase a text is to use a machine translation to translate the text to a pivot language, and then translating the text back to the original language, in the hope that these transformations will produce different sentences that are paraphrases. Duboue and Chu-Carroll (2006) investigated the usefulness of paraphrasing in question answering system by rephrasing the question. This was done by using multiple MT to translate a question to a pivot language and then translating it back to the original language. The best paraphrase is then selected. The approach however will suffer from error in translating back-and-forth.

3.3 Template Approaches

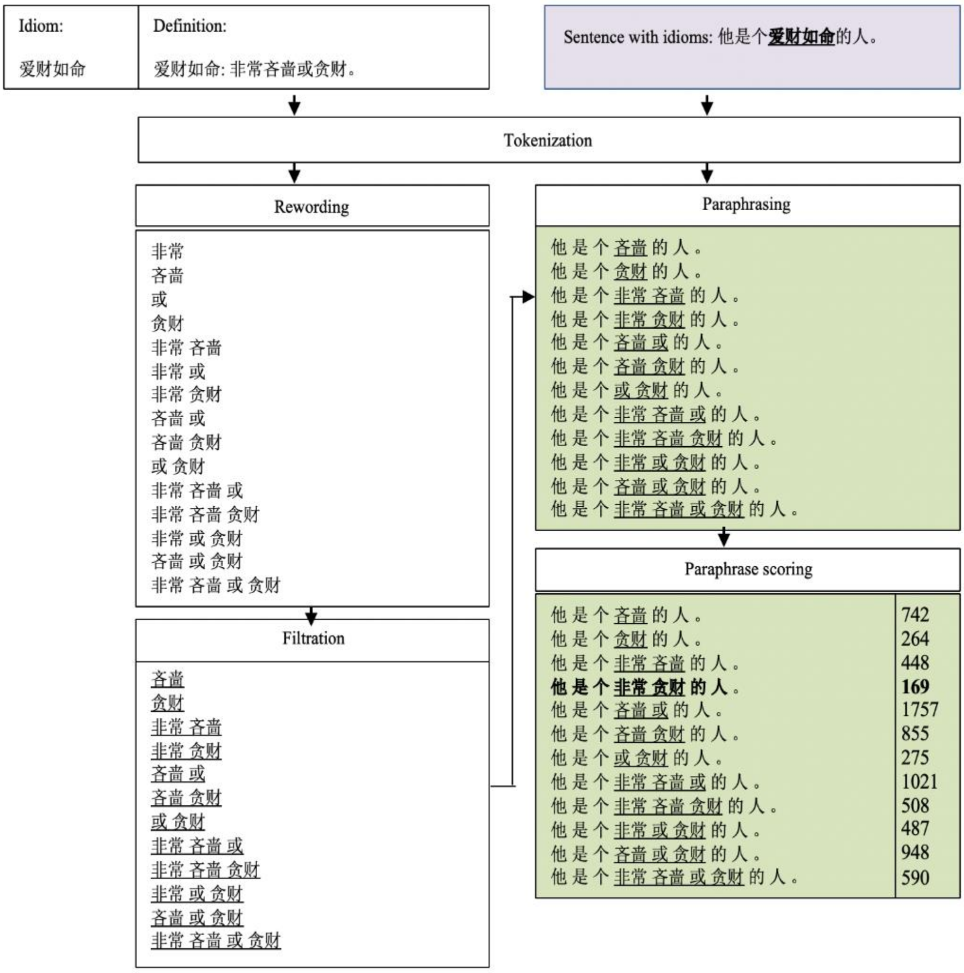
It is possible to derive templates for widely used sentences that describe certain thing in different ways. Ramono et al. (2006) found that 93% of the interacting protein pairs can be potentially identified using the template-based approach. Different templates such as “X interact with Y”, “X bind to Y”, “X Y complex”, and “interaction between X and Y” were used to describe the interactions of protein pairs X and Y that has the same meaning. Given two or more templates, new paraphrases can be generated given any X and Y values. The template can be created manually by experts, or an automatic approach can be used to derive possible templates. Szpektor et al. (2004) and Islam (2016) propose unsupervised method that acquire entailment relations from the Internet. Algorithms that solves paraphrasing and textual entailment are very similar, as paraphrasing can be consider as bidirectional textual entailment (Androutsopoulos and Malakasiotis, 2010).

4. Materials and Methods

Paraphrases consist of idioms are very limited. Nevertheless, idiom dictionaries containing the definitions of the idioms that serve as a useful resource for acquiring paraphrases. The definitions describe the meanings of the idiom and they may contain synonyms for the idiom. In this section, we describe our work to extract the definitions of Chinese idioms and using them for paraphrasing. Figure 1 shows our proposed approach to generate paraphrase for an idiom. Initially, paraphrase candidates for all idioms are extracted from dictionaries. The paraphrase candidates will go through tokenization, rewording, and filtration. Refer to the steps in Figure 1 on the left. Given an unknown sentence that contains an idiom, all possible paraphrase candidates for that idiom in the given context will beevaluated. The sentences will be scored to select the most suitable paraphrase for the idiom. See steps on the right in Figure 1.

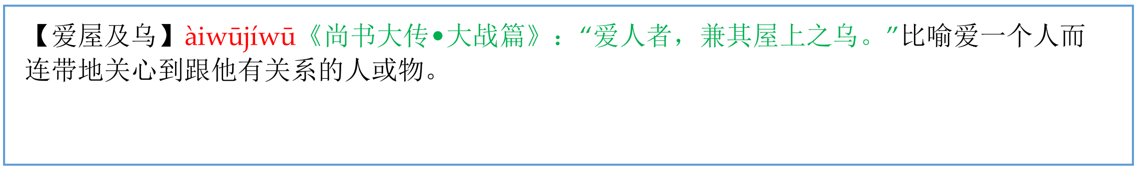
4.1 Extracting definitions from Chinese idiom dictionary

There are many Chinese idiom dictionaries in the market, most of them follow a similar compilation rule. A typical Chinese idiom entry includes the pronunciation, source, definition, example sentences, synonymous idioms, and other relevant information. Most idiom dictionaries will contain the pronunciations of the idiom in pinyin, and also source and definition of the idiom. The pinyin can be identified easily since it normally follows after the word or in a certain order. An example of the idiom 爱屋及乌 is provided in Figure 2 below. A chengyu dictionary may also contain information about the original source of the idioms, which is the document and sentence in Classical Chinese where the idiom was derived from.



他 是 个 吝啬 的 人 。	742
他 是 个 贪财 的 人 。	264
他 是 个 非常 吝啬 的 人 。	448
他 是 个 非常 贪财 的 人 。	**169**
他 是 个 吝啬 或 的 人 。	1757
他 是 个 吝啬 贪财 的 人 。	855
他 是 个 或 贪财 的 人 。	275
他 是 个 非常 吝啬 或 的 人 。	1021
他 是 个 非常 吝啬 贪财 的 人 。	508
他 是 个 非常 或 贪财 的 人 。	487
他 是 个 吝啬 或 贪财 的 人 。	948
他 是 个 非常 吝啬 或 贪财 的 人 。	590

Figure 1. Reordering the words in paraphrase to create new paraphrases.



【爱屋及乌】àiwūjíwū《尚书大传•大战篇》：“爱人者，兼其屋上之乌。”比喻爱一个人而连带地关心到跟他有关系的人或物。

Figure 2. The chengyu 爱屋及乌 in the Contemporary Chinese Dictionary

Most of the definitions of the idioms in the dictionary start with a prompt word. The idioms that are adjectives have the prompt word ‘形容’, which means “to describe”. The idioms that are used to express something metaphorically has the prompt word ‘比喻’, means ‘to metaphorize’. For example, the idiom ‘稳如泰山’ (English: as steady as the Taishan mountain) has this prompt word. The idioms where the definition starts with the prompt ‘指’ is normally used for idioms where the meaning of such idioms can often be inferred or extended literally from the words. These prompt words and other similar prompt words have to be deleted from the definitions. Below, we list some of these prompt words.

- 也形容 (also describe), 多形容 (often describe), 后多形容 (later often describe), 现形容 (now describe), 原形容 (originally describe), 旧时形容 (previously describe)
- 也比喻 (also metaphorize), 后比喻 (later metaphorize), 原比喻 (originally metaphorize), 现比喻 (now metaphorize), 现多比喻 (now often methaphorize)
- 原指 (originally means), 后指 (later means), 现指 (now means), 也指 (also means), 现也指 (now also means), 有时也指 (sometimes also means)

4.2 Tokenization

Chinese words and punctuations in a sentence are connected without any space. Thus, a tokenizer is required to segment the text to tokens.

4.3 Rewording and filtering

There is a possibility that some words in the definition has to be deleted when the definition substitutes an idiom in the sentence, because the definition may contain similar expressions already exist in the sentence. Consider the paraphrase "... is a lecturer" and "... teaches at a university". If there is a new sentence "Dr. Koehn **is a lecturer** at Johns Hopkins University.", the sentence can be paraphrased as "Dr. Koehn **teaches at a university** at Johns Hopkins University.". Even though the sentence is correct in this case, it is more appropriate to paraphrase the sentence as "Dr. Koehn **teaches** at John Hopkins University.". Our propose approach achieves this by deleting different combination of words in the definition to create different possible paraphrase candidates. If there are n words in the definition, there are 2^{n-1} paraphrase candidates that will produced. Since the words are repeated, the time complexity of the approach can be reduced using dynamic programming to $O(n^3)$. The syntax of the paraphrase candidates will then be analysed using dependency tree to filter out the text with an invalid syntax. This is done by analysing the head- dependent relationship in a dependency grammar. A dependency tree consists of internal nodes that are content words (e.g. noun, verb, adjective etc.) and leaves that are function words (e.g. preposition, auxiliary verb, conjunction etc.) or punctuations. A function word can only be used as the dependent. If in a sentence, the head is deleted, then the dependent that consists of function word must be deleted also, or else the syntax of the sentence produced will be invalid. Using this rule, some invalid paraphrase candidates are deleted.

4.4 Paraphrase selection

To find the best paraphrase candidates that best substitute the idiom for the context, a score will be calculated for each sentence. We propose to use perplexity for this purpose. The perplexity is the metric normally used for evaluating a language model given some test sentences. The lower the perplexity produced by the language model, the better it is in modeling the text. The equation [1] calculates the perplexity of a sentence given a trigram language model.

$$PP_{trigram}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1}, w_{i-2})}} \qquad [1]$$

$$W' = argmin(PP_{trigram}(W)) \qquad [2]$$

where$PP_{trigram}$ is the perplexity for a trigram language model given the sentence W (W= w_1, w_2, ..., w_N). P(w_i | w_{i-1}, w_{i-2}) is the conditional probability of word i given word i-1 and word i-2 that can be calculated as count(w_i, w_{i-1}, w_{i-2})/count(w_{i-1}, w_{i-2}). N is the number of words in the sentence. For paraphrasing, the perplexity is used to as a metric to select the most suitable paraphrase candidates for the given context. *The sentence with the lowest perplexity score is selected.For example, for the sentence W_1, "他是个吝啬的人。"*,

$$PP_{trigram}(W_1) = \sqrt[6]{1/P(\text{他}|<s>,<s>).P(\text{是}|<s>,\text{他}).P(\text{个}|\text{他，是})\dots P(\text{人}|\text{吝啬},\text{的}).P(\text{。}|\text{的，人})}$$

$PP_{trigram}$ for all the sentences will be calculated. The sentence with the lowest perplexity is selected.

5. Results and Discussion

We collected idiom definitions and sentences to evaluate the proposed approach. 500 idioms were randomly selected from two Chinese idiom dictionaries, 现代汉语词典(Chinese Academy of Sciences, 2002) and an online Chinese idiom dictionary,成语大全(http://tools.2345.com/chengyu). The idioms and the definitions were extracted and processed as describe in Section 4. For testing, we extracted sentences containing Chinese idioms from the idiom dictionary, 500 Common Chinese Idioms: An Annotated Frequency Dictionary (Jiao et al., 2011). 930 example sentences containing idioms were extracted from the dictionary. However, only 454 of the sentences were evaluated for paraphrasing that we have the definitions of the idioms. A Chinese native speaker was employed to manually paraphrase the idioms in the sentences.

We use THULAC for tokenization and LTP parser for creating dependency trees to filter out invalid sentences. For calculating the perplexity, we need to have a Chinese n-gram language model. We used more than 1 GB of Chinese Wikipedia text to create a 5-grams language model using SRILM (Stolcke, 2011). The 5-gram backoff language model was created using Kneser-Ney discounting algorithm and consist of 20 thousand vocabularies.

We compare our proposed approach against a baseline approach where the paraphrase for the idiom was selected from one of the definitions (without any editing of the definitions described in Section 4.3). If there are more than one definition, the definition with the lowest perplexity when substituting the sentence will be selected as the paraphrase. To evaluate the performance of different approaches, we used a similarity metric that allows us to compare the paraphrases and the references, the BLEU metric to compare the performance of the approaches. BLEU is a similarity metric widely used in machine translation, and the BLEU script from Moses SMT was used (Papineni, 2002). The higher the similarity between the translation sentences and the reference sentences, the higher the BLEU. Eq [3] is the formula to calculate BLEU.

$$BLEU = minimum\left(1, \frac{length\ of\ hypothesis}{length\ of\ reference}\right).\left(\prod_{n=1}^{4} precision_n\right)^{1/4} \qquad [3]$$

where length of hypothesis is the number of words in the paraphrase sentences, length of reference is the number of words in the reference sentences, and n is the n-gram order. We compared our proposed approach with a baseline approach that we describe above. Table 2 presents the results that we obtained.

Table 1. Paraphrasing result

Approaches	BLEU (%)
Baseline (without step filtering & rewording)	66.34
Proposed paraphrasing (LM-ngram)	75.69

The result show that our proposed method has a higher BLEU of 75.69%, compared to the baseline approach, which means the proposed approach produces paraphrases that match the reference more compared to the baseline approach.

6. Conclusion

In this paper, we describe our work in paraphrasing Chinese idioms. Our proposed approach uses the definitions from the idiom dictionaries as possible paraphrase candidates for the idioms. In addition, the definitions are edited to find possible better paraphrase candidates for the given context. The proposed paraphrasing approach rewords the definitions and produces new paraphrase candidates. The approach then scores the paraphrase candidates in the sentence by using perplexity. It is compared against the baseline approach that select the most suitable definition without any modification. The result shows that the proposed approach outperformed the baseline approach. Nevertheless, the n-gram language model limitation to n (, where normally 3<=n<=5) context words means it is unable to solve sentence with a long context. For future work, the study of RNN language model for the proposed approach will be interesting. In subsequent future works, studies will be carried out to evaluate the performance of the proposed approach on idioms from other languages.

7. Acknowledgment

The authors would like to express appreciation for the support of the Universiti Sains Malaysia [Project Number = 304.PKOMP.6316283].

References

1. Androutsopoulos, I. and P. Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research, 38, 135-187.
2. Barzilay, R. and K.R. McKeown. 2001. Extracting paraphrases from a parallel corpus. Proceedings of ACL/EACL, Toulouse, 50-57.
3. Chinese Academy of Sciences, ed. 2002. 现代汉语词典 [Contemporary Chinese Dictionary], Beijing: Foreign Language Teaching and Research Press.
4. Dolan, W. B. and C. Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. Proceedings of the Third International Workshop on Paraphrasing, Jeju Island, 9-16.
5. Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. Cambridge, MIT Press.
6. Ganitkevitch, J., B. Van Durme and C. Callison-Burch. 2013. Ppdb: The paraphrase database. Proceedings of the ACL: Human Language Technologies, Atlanta, 758–764.
7. Islam, R., Ghani, A.B.A., Kusuma, B., Theseira, B.B. (2016). Education and human capital effect on Malaysian economic growth. International Journal of Economics and Financial Issues, 6 (4), pp. 1722-1728.
8. Jiao, L., C.C. Kubler, W. Zhang. 2011. 500 Common Chinese Idioms: An annotated Frequency Dictionary, Routledge, Abingdon.

9. Luong, M.-T. 2016. Neural machine translation, Stanford University, Thesis.
10. Michal, Z., M. Junczys-Dowmunt and B. Pouliquen. 2016. The United Nations parallel corpus v1. 0. Proceedings of LREC, Portorož.
11. Mihalcea, R., C. Corley and C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. Proceedings of the Artificial intelligence, 775–780.
12. Papineni, K., S. Roukos, T. Ward and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of ACL, Philadelphia, 311-318.
13. Pedersen, T., S. Patwardhan and J. Michelizzi. 2004. WordNet::Similarity: measuring the relatedness of concepts. Proceedings of HLT-NAACL, Boston, 38-41.
14. Pershina, M., Y. He and R. Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. Proceedings of EMNLP, Lisbon, 76-82.
15. Ramono, L., M. Kouylekov and I. Szpektor. 2006. Investigating a generic paraphrase-based approach for relation extraction. Conference of the ACL, Trento, 409-416.
16. Stolcke, A., J. Zheng, W. Wang and V. Abrash. 2011. SRILM at sixteen: Update and outlook. Proceedings of ASRU Workshop, Hawaii.
17. Szpektor, I., H. Tanev, I. Dagan, and B. Coppola. 2004. Scaling web-based acquisition of entailment relations. Proceedings of EMNLP, Doha, 49-56.