A Consideration on the Methodology for Evaluating Large-scale Paraphrase Lexicons

ATSUSHI FUJITA[1,a)]

Abstract: Aiming at creating paraphrase lexicons that ensure good coverage of the target classes of paraphrases along with a low proportion of incorrect information, in the last decade, researchers have proposed methods for extracting sub-sentential paraphrases from various types of corpora. Once a paraphrase lexicon is created, then the ensuing issue is how to measure its quality. This is typically performed through a substitution test: each of sampled pairs of expressions is judged whether it is a correct paraphrase pair or not by evaluating grammaticality and meaning equivalence of the expressions in actual sentences. In this paper, we describe the issues in evaluating paraphrase lexicons. Then, focusing on a widely-used evaluation scheme, i.e., substitution test for samples, we propose three extensions designed for obtaining a more consistent human judgments: (i) classification-based evaluation criteria, (ii) two-step unit-wise evaluation procedure, and (iii) re-evaluation of disagreed examples. Through an evaluation experiment, we have confirmed at least the third extension contributes to improve the inter-evaluator agreement ratio.

Keywords: Paraphrase, Knowledge acquisition, Evaluation, Inter-evaluator agreement

1. Introduction

One of the characteristics human languages have is that the same semantic content can be expressed with several different linguistic expressions, i.e., **paraphrases**. For instance, three sentences in example (1) are paraphrases of each other.

(1) a. The brothers *look like* each other.
b. The brothers *resemble* each other.
c. The brothers *are similar to* each other.

The substituted phrases "*look like*," "*resemble*," and "*are similar to*" can also be regarded as (sub-sentential) paraphrases of each other. Dealing with paraphrases is one of the common issues in a broad range of natural language processing (NLP) tasks. In other words, computational technologies that recognize and/or generate paraphrases robustly and accurately have a great potential to improve existing NLP applications, such as information retrieval (including Web search), machine translation, question answering, and text mining.

The notion of paraphrase covers diverse phenomena including lexical substitutions, such as those shown in example (1), syntactic transformation, and discourse-level reconstruction. Among them, to automate word- or phrase-level paraphrases that are heavily depending on each lexical item, a knowledge-base about words and phrases that have (approximately) same meaning, i.e., **paraphrase lexicon**, is the most essential resource. For instance, without knowing that "*look like*," "*resemble*," and "*are similar to*" have same meaning, computers (or even humans) cannot realize that the sentences in example (1) convey the same meaning.

In the last decade, creating paraphrase lexicons through extracting sub-sentential paraphrases from corpora has been drawing the attention of many researchers [12, 15, 1]. Typically, automatically acquired paraphrases are represented with pairs of words/phrases, such as ("*look like*", "*resemble*"). The challenge in acquiring paraphrases is to ensure good coverage of the targeted classes of paraphrases along with a low proportion of incorrect pairs.

Once a paraphrase lexicon is created, then the ensuing issue is how to measure its quality. As an intrinsic evaluation, human evaluations for sampled knowledge have been preferably conducted, because paraphrase lexicons that are automatically created tend to be too large for exhaustive evaluation. Since the work in [19], substitution-based evaluation has been getting popular as a way for judging whether each sampled paraphrase pair has approximately same meaning. For instance, a pair of phrases ("*look like*", "*resemble*") is judged through substituting them in actual sentences, such as shown in example (1). An improved evaluation criterion has been proposed in [5], where each substitution of paraphrase candidates is evaluated from two viewpoints, i.e., **grammaticality** and **meaning equivalence**.

Aiming to achieve highly consistent judgments from human evaluators in a widely-used evaluation scheme [5], this paper proposes three extensions. First, classification-based evaluation criteria are proposed instead of the numeric scoring. We created decision trees for guiding evaluators judging grammaticality and meaning equivalence of each word/phrase substitution. Second, a more controlled procedure is presented to explicitly distinguish the two evaluation viewpoints, i.e., grammaticality and meaning equivalence. Finally, a re-evaluation step is introduced. Each

[1] Faculty of Systems Information Science, Future University Hakodate
[a)] fujita@fun.ac.jp

evaluator is asked to reconsider some parts of her/his initial results to reduce disagreement between other evaluators.

Unfortunately, it is hard to justify the first two extensions, because two different evaluation schemes cannot coexist. Using one by one is also unreliable: if one has worked with an evaluation scheme, working with another scheme is biased to a certain degree. Depending on the task, asking another group of human evaluators may be an option; however, as diverse inter-evaluator agreement ratios in our experiment indicate, difficulty of evaluation is heavily depending on evaluators' proficiency in the language of interest (and certainly data). Thus, we mainly gauge the impact of the last extension, fixing evaluators and data sets.

The rest of this paper is organized as follows. First, existing methods for evaluating paraphrase lexicons are reviewed and remaining issues are summarized in Section 2. Then, we focus on the substitution-based evaluation. Section 3 describes the two viewpoints of evaluation and Section 4 introduces our extensions. Sections 5 and 6 are devoted to explain our evaluation experiment. Finally, Section 7 concludes this paper.

2. Existing methods and remaining issues

In this paper, we assume paraphrase lexicons that comprise pairs of actual expressions (words and phrases) or pairs of phrase patterns, acknowledging that standardization of representation of paraphrases is also an issue in the community. In fact, most of the previous work on the automatic paraphrase acquisition produced pairs of such expressions [3, 14, 2, 16, 4, 9, 5, 21, 13, 11, 10, 18, 20, 17].

First issue is how to take an appropriate set of samples. As mentioned above, paraphrase lexicons that are automatically created tend to be too large for exhaustive evaluation. Therefore, manual intrinsic evaluation can be performed only for samples. Preparing a data set independently from the acquisition process is an alternative way. Many of previous work have evaluated samples randomly taken from the entire lexicon. For the lexicons in which each pair of paraphrases is associated with some confidence/reliability scores, taking samples from a certain volume of the most reliable part of the lexicon is a reasonable option. For instance, Hashimoto et al. [11] and Yan et al. [20] demonstrated the trade-off between the volume and reliability scores. Such scores can be used for ranking multiple candidates for the same phrase. Therefore, some researchers studied the goodness of scoring/ranking functions [9, 6, 18, 17]. However, nothing can guarantee that samples obtained by such a way properly represent the entire lexicons.

Second issue is on the range of context considered in the evaluation. Although paraphrase lexicons tend to be generated relying on contextual information to some degree, if the lexicons are regarded as kinds of multi-purpose lexical resources, evaluating them separately from context makes sense. However, when applying them to a particular NLP task that refers to some context, the appropriateness of each paraphrase in that context has to be taken into account. Szpektor et al. [19] have revealed that providing evaluators with some contextual information improved the consistency of judgments. These are the major reasons why substitution-based evaluation has been getting popular as a way for judging each pair of paraphrases. Although one can evaluate a paraphrase comprising a single word replacement considering a whole document, it is excessive. Typically, moderate length of single sentences are used.

Evaluation criterion is indispensable for regulating human judgments. While the term *paraphrase* primarily indicates equivalence of meaning, given that an evaluation is performed considering the context, assessing whether the grammaticality suffers or not is also important. It is therefore straightforward to evaluate each example from these two viewpoints. Callison-burch [5] conducted an evaluation experiment considering there two viewpoints, employing a 5-point scale for each. However, grammaticality and meaning equivalence are confusing concepts. In our preliminary evaluation experiments following the same 5-point scales, some evaluators have given inconsistent scores, mainly mixing up the two concepts. This would be one of the reasons why some recent work, such as [13], utilized a simplified version of the scales.

Last but not least, the quality of human evaluation perfectly depends on the expertise of evaluators. None of the previous work has achieved the perfect inter-evaluator agreement ratio, i.e., 1.0. It is partly because of the insufficiency of the evaluation criteria and the difficulty of the task, but we guess evaluator's proficiency in the language of interest is also essential. Human evaluation is basically expensive and time-consuming. However, utilizing anonymous workers on the Web through so-called crowdsourcing platforms, such as Amazon Mechanical Turk (AMT)[*1], has been getting popular in a wide range of NLP tasks. AMT has been used for ranking multiple candidates for the same phrase [6]. The remaining issue is to establish a standard for managing diverse expertise of workers and sufficient instruction for beginners to ensure evaluation consistency.

3. Grammaticality and meaning equivalence

Generating paraphrase sentences by word/phrase substitution involves two different tasks: (re-)generating sentence and preserving meaning. It is therefore straightforward to evaluate examples in terms of **grammaticality** and **meaning equivalence** separately [5]. Let us now consider these two concepts.

Grammaticality: whether the paraphrased sentence is grammatical

Meaning: whether the meaning of the original sentence is properly retained by the paraphrased sentence

Let's see several examples to better understand the distinction between these two concepts. See example (2)[*2]. The paraphrased sentence has no grammatical problem. The substituted phrase "*environmental issues*," however, conveys different meaning from the original phrase "*global economy*." One may recognize that the original and the substituted phrases share some meaning, i.e., both are a kind of social issue.

(2) s. The leaders discussed the *global economy*.
    t. The leaders discussed the *environmental issues*.

See another example below. Grammaticality of the para-

[*1] https://www.mturk.com/

[*2] Throughout this paper, original and paraphrased sentences are labeled "s" and "t," respectively.

Table 1 Classification results by the author.

Example	Grammaticality	Meaning equivalence
(1)	Perfect	Equivalent
(2)	Perfect	Significantly Different
(3)	Irredeemable	Equivalent
(4)	Perfect	Significantly Different

phrased sentence suffers. One may be able to correct it without referring to the original sentence, while one cannot do so with the other examples. Meaning equivalence is evaluated irrespective of grammaticality. If one notices no additional information nor loss of information, even if there is a serious grammatical problem, the meanings of two sentences are evaluated to be equivalent.

(3) s. I like to be *30 years* old.
t. I like to be *age of 30* old.

Consider one more example which requires very careful judgment. Given a pair of the original phrase "*a movement against racism*" and its paraphrase "*an anti-racism movement*", one may recognize that they are semantically equivalent. However, substitution of these phrases within the sentence below collapses a coordination of two nominal elements "*racism* and fascism" and consequently changes the meaning of the original sentence. Interestingly, the paraphrase does not have any grammatical problem despite the collapse of the coordination in the original sentence.

(4) s. They expressed support for *a movement against racism* and fascism in Athens.
t. They expressed support for *an anti-racism movement* and fascism in Athens.

4. Evaluation procedure

This section explains the following three extensions that we introduce on top of the previously proposed evaluation scheme.

- Classification-based evaluation criteria
- Unit-wise two-step evaluation procedure
- Re-evaluation of examples that had inconsistent judges

4.1 Classification-based evaluation criteria

Callison-burch [5] asked his evaluators to rate grammaticality and meaning equivalence along two 5-point scales: 5 means good and 1 means bad. Although what each score means is explained, numerical scores have a potential drawback: evaluators might give scores subjectively and/or intuitively without a careful consideration, as if rating products and movies. In fact, in our preliminary experiment performed in the same manner, we observed some evaluators who gave different scores to examples that have only the same types of errors. We should minimize this type of (intra-evaluator) inconsistency as much as possible.

Instead of the 5-point scales, we ask our human evaluators to label each example with one of the predefined categories. In other words, human evaluators are asked to perform a classification task. Moreover, aiming at even more consistent results, we provide evaluators with decision trees composed of atomic questions (see Appendices A.1 and A.2). Assigning a category label to an example corresponds to giving a combination of answers to the questions. Table 1 shows classification results for examples (1) to (4) made by the author.

4.2 Unit-wise two-step evaluation procedure

Aiming at making results further consistent, we provide evaluators with several paraphrases per source phrase token at the same time. We call each set of examples a **unit**. This approach also reduces the human labor spent for reading and understanding the original sentences repeatedly. Unlike the functionality of the current crowdsourcing platforms, such as AMT, our evaluation tool, which is tailored for this evaluation task, allows evaluators to postpone and revise their judgment for individual examples. Human evaluators are encouraged to reconsider examples that they have already judged to make results consistent as much as possible.

As already mentioned, grammaticality and meaning equivalence are confusing concepts. We attempt to minimize such confusion by controlling the order of evaluation as follows.

Step 1. Grammaticality first: In the first step, when a unit is specified by the evaluator, only paraphrased sentences are shown. Evaluators judge their respective grammaticality without seeing the original sentence. If one finds a paraphrased sentence ungrammatical in a way that is not caused by the substituted phrase marked italic and underline, she/he ignore it because it is inherited from the original sentence.

Step 2. Then meaning equivalence: Once the classification results for the grammaticality are submitted, the paraphrased sentences are again shown along with their original sentence. Evaluators judge to what extent the meaning of the original sentence is retained by each paraphrased sentence. This is a way of evaluating the meaning equivalence of substituted phrases per se as well.

4.3 Re-Evaluation of examples

While the unit-wise two-step procedure described above aims to improve **intra-evaluator consistency**, we also make an attempt to improve **inter-evaluator consistency** by introducing a re-evaluation phase.

The aim of re-evaluation is to solve disagreement between evaluators; however, we should avoid compromises where they simply meet in the middle. Furthermore, there might be examples that get the same but wrong labels from evaluators. To balance between the human labor and the chance of potential correction, we provide evaluators with two sets of examples after mixed and shuffled. The first set comprises examples that yield disagreement at the level of the following coarse-grained **binary classes** that we have determined on the basis of the work in [5] (see Appendices A.1 and A.2 for the content of questions).

Grammaticality: if Q1 is answered "Yes" and as a consequence the assigned label is either "Perfect" or "Awkward," the grammaticality of the example is regarded OK; otherwise NG.

Meaning equivalence: if Q1 is answered "Yes," Q2 is answered "No," and as a consequence the assigned label is either "Equivalent," "Missing Info.," "Additional Info.," or "Ignorable Change," the meaning equivalence of the example is regarded as OK; otherwise NG.

Another set comprises randomly sampled examples that get the same binary classes in the first evaluation phase.

Evaluators are asked to reconsider the labels of the provided examples that they have given in the first evaluation phase. They are instructed that (i) they do not necessarily have to change their initial decision, (ii) they are also encouraged to look at examples that are not marked for re-evaluation.

5. Experimental settings

To clarify to what extent the proposed extensions improve the evaluation results, we have conducted an experiment.

5.1 Data

In this experiment, we took three English paraphrase lexicons created by the following method proposed in [10]. A brief summary of the creation procedure of the lexicons is given below.

Step 1. Seed paraphrase acquisition: Seed paraphrases were extracted from bilingual corpora, regarding shared phrasal translations as the evidence of semantic equivalence of phrases [2]. From the English-French version of the Europarl Parallel Corpus (release 6)[*3], 4.0 million paraphrase pairs for 911 thousand unique phrases (P_{Raw}) were obtained. By applying several filters, 1.2 million pairs for 450 thousand unique phrases (P_{Seed}) were retained. While only pairs in P_{Seed} were used in the following step, pairs that were filtered out ($P_{Del} = P_{Raw} \backslash P_{Seed}$) were also used in the evaluation.

Step 2. Paraphrase pattern induction: From the seed paraphrases, paraphrase patterns, such as ("X system", "X apparatus"), were learned.

Step 3. Paraphrase instance acquisition: Using the learned patterns, a novel set of paraphrase pairs were harvested from monolingual non-parallel corpora: (i) the English Gigaword Corpus (Fifth Edition)[*4], (ii) the English side of the 10^9 French-English corpus[*5], and (iii) the English side of the above Europarl Parallel Corpus. As a result, 62 million paraphrase pairs for 20 million unique phrases (P_{Hvst}) were obtained.

We decided to show 5 different paraphrases in each unit. To generate such test units, the source-side of paraphrase pairs were restricted to those having at least 5 paraphrases in $P_{Del} \cup P_{Seed} \cup P_{Hvst}$. As a result, 37.6 million paraphrase pairs for 3.8 million unique phrases were retained.

Similarly to previous work [5, 10], we used news sentences. To be precise, the English part of WMT 2011 "newstest" data consisting of 3,002 unique sentences was used. To reduce the human labor for the evaluation, sentences were restricted to those with moderate length: 10-30 words, which we expected to provide sufficient but succinct context. As a result, 1,919 sentences were retained.

Test units were generated in the following manner:

Step 1. By applying the paraphrase lexicons P_{Del}, P_{Seed}, and P_{Hvst} to the test sentence set, 18,833 phrase tokens (11,955 unique phrases) in 1,911 sentences were paraphrased and 191,382 unique pairs of sentences were generated.

[*3] http://statmt.org/europarl/
[*4] LDC Catalog No. LDC2011T07
[*5] http://statmt.org/wmt10/training-giga-fren.tar

Table 2 Distribution of paraphrases.

Data set	n	P_{Raw}	P_{Seed}	P_{Hvst}	Note
Data 1	405	258	105	149	$\lvert P_{Raw} \cap P_{Hvst} \rvert = 2$
Data 2	595	76	22	519	

Step 2. For each source phrase token, 5 pairs of sentences (i.e., 5 paraphrases) were randomly selected.

Step 3. Data 1 is compiled to compare paraphrases for the same phrase token but from different lexicons. In other words, noisiness of the state-of-the-art resource, i.e., P_{Raw} ($= P_{Del} \cup P_{Seed}$) is evaluated[*6]. We extracted all units that contain at least one paraphrase from each of P_{Del}, P_{Seed}, and P_{Hvst}. As a result, 81 units were obtained.

Step 4. Another 119 units (henceforth, **Data 2**) were prepared by extracting those containing at least one paraphrase from P_{Hvst}. Because this set contains more examples from the extended resource, i.e., P_{Hvst}, its quality would be more accurately estimated.

Table 2 shows the distribution of paraphrases within two data sets. Two data sets, consisting of 200 units (1,000 examples) in total, were mixed and shuffled; then different subsets were distributed to human evaluators.

5.2 Evaluation

We recruited six evaluators who had completed translation studies at a university and passed a preliminary screening based on the English proficiency.

Each human evaluator was first asked to work on 100 units (each unit was judged by 3 different evaluators), following the procedure described in Section 4.2. After all six evaluators completed their work, the results were compared and the examples that needed to be re-evaluated (see Section 4.3) were given to be evaluators, along with their original labels (given by herself/himself). In average, 167 out of 500 examples required re-evaluating their grammaticality and 222 examples did so for their meaning equivalence. As described in Section 4.3, some examples that had consistent judgments at the binary level were also sent back to the evaluators. We randomly sampled such examples just 10% of the above, i.e., 17 and 22 examples for the grammaticality and meaning equivalence viewpoints, respectively.

6. Results

6.1 Agreement ratio

How consistently did the evaluators give labels to the examples? We calculated the agreement ratio at the two different levels. G_5 and M_6 refer the results based on the fine-grained classification: the numbers indicate those of the predefined categories. On the other hand, G_2 and M_2 indicate classification results based on the binary classes.

The agreement ratio of the entire data set is quantified by the Fleiss' κ [8], which accounts for not only the number of examples that actually get the same label but also the potential agreement that happens by chance. The results are summarized in Table 3. The table shows a visibly large improvement of the agreement ra-

[*6] Although the focus of this paper is the evaluation methodology, we also describe the precision of the examined paraphrase lexicons in Section 6.

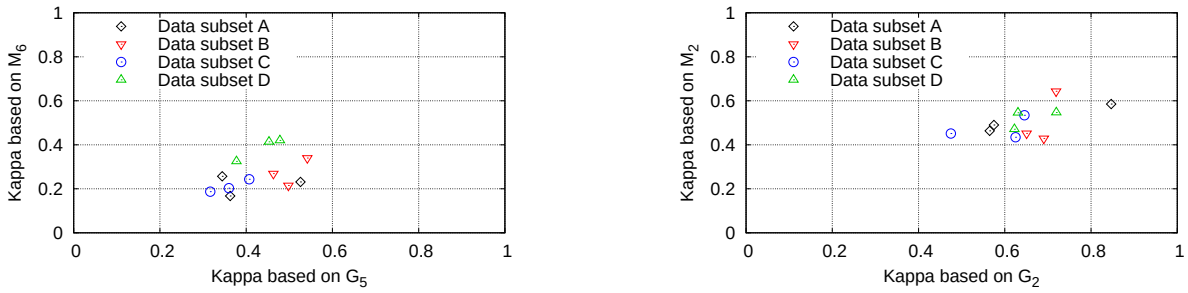

Fig. 1 Pairwise Cohen's κ values for the results **before** re-evaluation ($n = 250$ for each point).

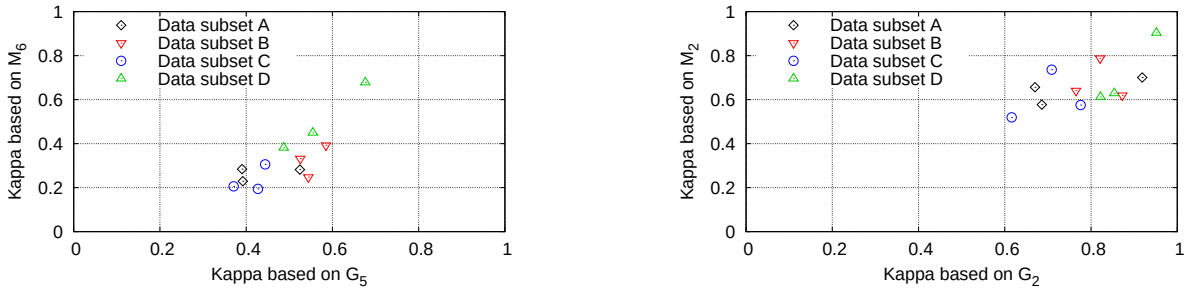

Fig. 2 Pairwise Cohen's κ values for the results **after** re-evaluation ($n = 250$ for each point).

Table 3 Fleiss' κ values ($n = 1{,}000$).

Phase	G_5	M_6	G_2	M_2
Before re-evaluation	0.427	0.268	0.648	0.500
After re-evaluation	0.494	0.325	0.789	0.659

Table 4 Precision of the evaluated examples in Data 1.

Paraphrase lexicon	n	G	M	Both
P_{Raw} (State-of-the-art)	258	0.46	0.83	0.38
P_{Del} ($= P_{Raw} \backslash P_{Seed}$)	153	0.33	0.81	0.23
P_{Seed} ($= P_{Raw} \backslash P_{Del}$)	105	0.65	0.87	0.59
P_{Hvst}	149	0.56	0.53	0.36

tio through re-evaluation. The final results for G_2 (0.789) and M_2 (0.659) can be considered substantial. Figures 1 and 2 elaborate the agreement ratio of individual pairs of evaluators calculated by Cohen's κ [7]. These figures also show that the agreement ratio is generally improved irrespective of the pair of evaluators and subset of the evaluation data. We also observed that the κ values were diverse depending on the pair of evaluators (and data set) and the evaluation of meaning equivalence is more difficult than judging grammaticality, given our evaluation criteria.

One may be interested in to what extent the results are superior or whether the method is comparable to the state-of-the-art evaluation schemes. However, it is hard to perform a fair comparison, because we could not employ the same evaluators and data nor replicate the experiment performed in the previous work without introducing any bias. We give several values below just for reference. Callison-burch [5] collected judgments for 1,391 examples from 2 evaluators and reported κ values, 0.33 for 5-point scales and 0.61 for its coarse-grained binary version. Kok and Brockett [13] performed an evaluation using a 3-point scale designed by combining and simplifying the original two 5-point scales and obtained 0.62 for the κ value.

Note that a high κ value does not necessarily indicate that the obtained labels are truly correct. Ensuring correctness is an open question in the community.

6.2 Estimated quality of paraphrase lexicons

Although the focus of this paper is the consistency of human evaluation, we also report on the estimated quality of the examined paraphrase lexicons. As a measure of the quality, we calculated precision of the evaluated examples: if the majority of evaluators (two or three) assigned a label corresponding to OK class in the binary decision, the example was regarded as correct.

Table 4 shows the precision of examples generated using the paraphrases within each lexicon. There is a clear performance gap between P_{Raw} and P_{Seed} in terms of the grammaticality (0.46 vs 0.65). This proves that the filters proposed in [10] have properly discarded paraphrases that hurt grammaticality (P_{Del}), keeping the high level of meaning equivalence evidenced by the shared translations. On the other hand, P_{Hvst} had lower precision than P_{Seed} for both grammaticality and meaning equivalence. An expected advantage of P_{Hvst} is that it can cover phrases and their paraphrases that are not directly obtainable from parallel corpus as P_{Seed}; however, phrases that are already covered with P_{Seed} may not necessarily need such an expansion.

Each paraphrase pair in P_{Seed} and P_{Hvst} is associated with a score indicating how likely the pair of expressions is to be paraphrases, which is estimated on the basis of so-called contextual similarity. We assumed that we could control the precision by varying the threshold for this score. However, our result showed that though our assumption above was roughly true, it was imperfect. Curves in Figure 3 show precision of P_{Seed} and P_{Hvst} in Data 1 and P_{Hvst} in Data 2. Generally, the higher the threshold is, the higher the precision is. However, meaning equivalence of P_{Seed} tested on Data 1 has a rather stable shape and the P_{Hvst} tested on Data 2 has a large drop of precision at a high threshold in both of grammaticality and meaning equivalence. An intuitive explanation of the former is that the meaning equivalence of paraphrases in P_{Seed} is highly guaranteed by the shared translations,

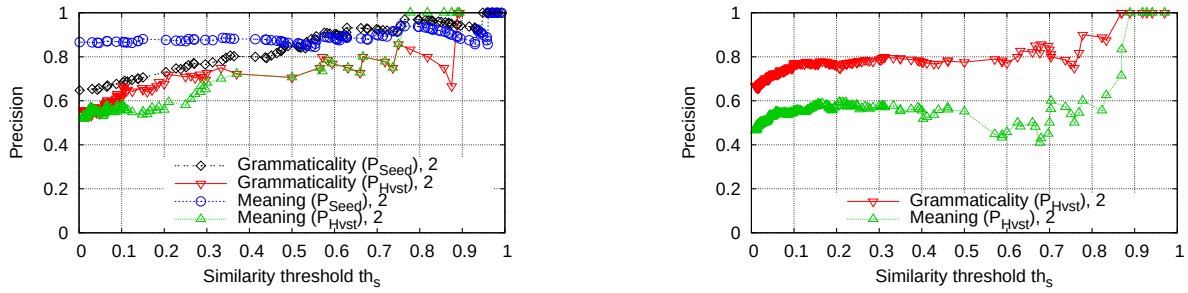

Fig. 3 Precision against threshold values (left: Data 1, right: Data 2).

so the contextual similarity is not further helpful. In contrast, the low precision of P_{Hvst} may be caused by several reasons, such as paraphrase pattern induction, paraphrase instance acquisition, and the estimation of the score.

Note that the low precision in total is because of the naive method for generating paraphrased sentences, i.e., phrase replacement. It is promising that other features, such as statistical language model scores, boost precision [5].

7. Conclusion

In this paper, we first described the issues in evaluating gigantic paraphrase lexicons that are automatically created. Then, on top of an existing evaluation scheme, we introduced three extensions aiming to improve consistency of evaluation, i.e., we elaborated an evaluation criterion for both grammaticality and meaning equivalence, two-step unit-wise evaluation procedure, and performing re-evaluation for examples that had inconsistent results. It is hard to quantify the effectiveness of the first two features. In contrast, we confirmed that introducing a re-evaluation phase always improved the inter-annotator agreement. Our future work includes a further investigation into the evaluation methodology to reduce human labor, while retaining high data quality.

Acknowledgments

The author is deeply grateful to Elizabeth Marshman and Pierre Isabelle for their kind support and valuable comments. This work was done when the author was at the National Research Council Canada as a JSPS (the Japan Society for the Promotion of Science) Postdoctoral Fellow for Research Abroad.

References

[1] Androutsopoulos, I. and Malakasiotis, P.: A Survey of Paraphrasing and Textual Entailment Methods, *Journal of Artificial Intelligence Research*, Vol. 38, pp. 135–187 (2010).
[2] Bannard, C. and Callison-Burch, C.: Paraphrasing with Bilingual Parallel Corpora, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 597–604 (2005).
[3] Barzilay, R. and McKeown, K. R.: Extracting Paraphrases from a Parallel Corpus, *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 50–57 (2001).
[4] Bhagat, R. and Ravichandran, D.: Large Scale Acquisition of Paraphrases for Learning Surface Patterns, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 161–170 (2008).
[5] Callison-Burch, C.: Syntactic Constraints on Paraphrases Extracted from Parallel Corpora, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 196–205 (2008).
[6] Chan, T. P., Callison-Burch, C. and Durme, B. V.: Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity, *Proceedings of the Workshop on Geometrical Models of Natual Language Semantics* (*GEMS*), pp. 33–42 (2011).
[7] Cohen, J.: A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46 (1960).
[8] Fleiss, J. F.: Measuring Nominal Scale Agreement among Many Raters, *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382 (1971).
[9] Fujita, A. and Sato, S.: A Probabilistic Model for Measuring Grammaticality and Similarity of Automatically Generated Paraphrases of Predicate Phrases, *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*), pp. 225–232 (2008).
[10] Fujita, A., Isabelle, P. and Kuhn, R.: Enlarging Paraphrase Collections through Generalization and Instantiation, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL*), pp. 631–642 (2012).
[11] Hashimoto, C., Torisawa, K., De Saeger, S., Kazama, J. and Kurohashi, S.: Extracting Paraphrases from Definition Sentences on the Web, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 1087–1097 (2011).
[12] Inui, K. and Fujita, A.: A Survey on Paraphrase Generation and Recognition, *Journal of Natural Language Processing*, Vol. 11, No. 5, pp. 151–198 (2004). (in Japanese).
[13] Kok, S. and Brockett, C.: Hitting the Right Paraphrases in Good Time, *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pp. 145–153 (2010).
[14] Lin, D. and Pantel, P.: Discovery of Inference Rules for Question Answering, *Natural Language Engineering*, Vol. 7, No. 4, pp. 343–360 (2001).
[15] Madnani, N. and Dorr, B. J.: Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods, *Computational Linguistics*, Vol. 36, No. 3, pp. 341–387 (2010).
[16] Paşca, M. and Dienes, P.: Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web, *Proceedings of the 2nd International Joint Conference on Natural Language Processing* (*IJCNLP*), pp. 119–130 (2005).
[17] Razmara, M., Siahbani, M., Haffari, C. and Sarkar, A.: Graph Propagation for Paraphrasing Out-of-Vocabulary Words in Statistical Machine Translation, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 1105–1115 (2013).
[18] Suzuki, Y., Sasano, R., Takamura, H. and Okumura, M.: *Rimu-ru*, *Doya-ru*, *Poji-ru*, *Pafe-ru*: Acquisition Paraphrases and Etymologies of Katakana Verbs from Web Corpora, *Information Processing Society of Japan SIG Notes, NL-209-8*, pp. 1–7 (2012). (in Japanese).
[19] Szpektor, I., Shnarch, E. and Dagan, I.: Instance-based Evaluation of Entailment Rule Acquisition, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 456–463 (2007).
[20] Yan, Y., Hashimoto, C., Torisawa, K., Kawai, T., Kazama, J. and De Saeger, S.: Minimally Supervised Method for Multilingual Paraphrase Extraction from Definition Sentences on the Web, *Proceedings of Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pp. 63–73 (2013).
[21] Zhao, S., Wang, H., Liu, T. and Li, S.: Extracting Paraphrase Patterns from Bilingual Parallel Corpora, *Natural Language Engineering*, Vol. 15, No. 4, pp. 503–526 (2009).

Appendix

Appendices A and B show decision trees and component questions presented to human evaluators. They were used for evaluating grammaticality and meaning equivalence, respectively.

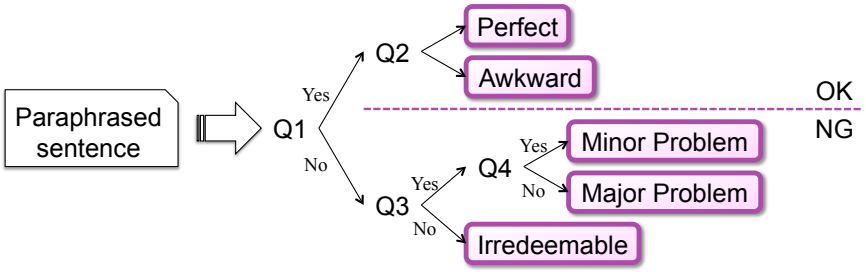

Fig. A·1 Decision tree for evaluating grammaticality.

A.1 Grammaticality: Is the paraphrase grammatical?

Given the paraphrase, answer the following questions without seeing the original sentence.

Q1: Is it grammatical?

Yes: answer Q2

- apart from whether it is true or not: e.g., "I saw a unicorn yesterday."
- apart from whether it is nonsensical or not: e.g., "Colorless green ideas sleep furiously."

No: answer Q3

Q2: Is it perfectly grammatical or something awkward?

Perfect: label it "Perfect"

Awkward: label it "Awkward"

- Strange collocation: e.g., "Individual members are equipped with *strong computer* systems."
- Fail to form a contrast: e.g., "Eleven **men** and *three* ***workers*** were arrested."
- Stylistically inconsistent: e.g., "In each category, this award totals *10* ***m*** Swedish krona (approximately 25 **million** CZK)."
- etc.

Q3: Is the grammatical error correctable?

Yes: answer Q4

No: label it "Irredeemable"

Q4: Is the grammatical error corrected with only one edit, such as the followings?

- Deletion of unnecessary word: e.g., "*thirty years* ***old*** **old**"
- Correction, deletion, or addition of determiner: e.g., "**a** ***ambitious*** *level* of advantage"
- Correction of hyphenation error:
  e.g., "The Bank of England replies to concerns by lending 10 billion pounds for ***5-weeks***."
- Correction of mismatch between present and past:
  e.g., "The *commission* ***report*** that BSkyB's stake **thwarted** competition and **allowed** it unfair influence over ITV."
- Correction of agreement error (between subject and verb, between a plural noun and singular determiner, etc.):
  e.g., "The *commercial* ***results*** of the US **feeds** optimism."
- etc.

Yes: label it "Minor Problem"

No, more than that: label it "Major Problem"

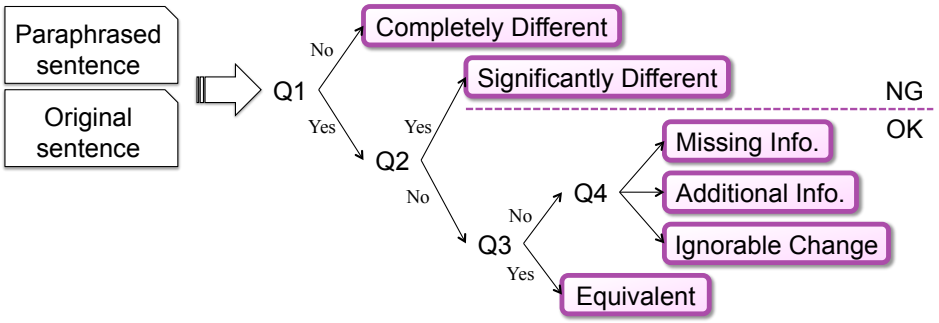

Fig. A·2 Decision tree for evaluating equivalence of meaning.

A.2 Meaning: Does the paraphrase preserve the meaning of the original sentence?

Given a pair of paraphrase and its original sentence, answer the following questions.

Q1: Do the two phrases share some meaning in this context?

Yes: answer Q2

No: label it "Completely Different"

Q2: Does the paraphrase convey a meaning significantly different from the original sentence in this context?

Yes: label it "Significantly Different"

- Different: e.g., "He waited for *two years*." ⇒ "He waited for *three years*."
- Different:
  e.g., "Gaudi designed a *central heating* system in the house."
  ⇒ "Gaudi designed a *first heating* system in the house."
- Narrowing the area is critical:
  e.g., "The leaders discussed the *global economy*."
  ⇒ "The leaders discussed *the economic issues in Europe*."
- Broadening the area is critical:
  e.g., "The leaders discussed the *economic issues in Europe*."
  ⇒ "The leaders discussed the *global economy*."
- etc.

No, nothing is changed or there are only ignorable changes: answer Q3

Q3: Are the meaning that two sentences convey perfectly equivalent?

Yes: label it "Equivalent"

No: answer Q4

Q4: Is the (slight) difference between two sentences?

Loss: label it "Missing Info."
e.g., "The baby boom crested *around 1957*." ⇒ "The baby boom crested *in the late 1950s*."

Addition: label it "Additional Info."
e.g., "*Twelve million* people were affected in the crash." ⇒ "*12.00 million* people were affected in the crash."

Something else including both loss and addition: label it "Ignorable Change"