Effective Paraphrase Expansion in Addressing Lexical Variability

Vasily Konovalov, Meni Adler, Ido Dagan
Bar-Ilan University
Ramat-Gan, Israel
vaskonov@yahoo.com, meni.adler@gmail.com, dagan@cs.biu.ac.il

***Abstract*—In this paper, we investigate the contribution of automatically generated translation-based paraphrases to address lexical variability with application in dialogue systems. We compare the proposed methods with the state of the art approach. Furthermore, we define the desired criteria for the pivot language to generate the paraphrases, and find that the performance of paraphrase expansion correlates well with the averaged smoothed BLEU measure. The results suggest that: (1) The paraphrase expansion leads to better performance than the bag of words baseline. (2) The gain in performance often comes from a few most effective pivot languages. (3) The differences between machine translation engines are not reflected in empirical evaluation. By using the most effective pivot languages we can save the expenses to generate additional paraphrases, and as a result, to save the resources to train a classification model.**

I. Introduction

One of the most interesting properties of natural language is *lexical variability* - representing the same idea in different ways; for example, the concept of *salary* could be expressed as *wage*, *pay*, *earnings*, etc. Robust natural language applications should be able to deal with lexical variability.

Due to lexical variability, a concept of 'Agreement' in Switchboard telephone conversation corpus [1] can be expressed as *"that would be a real good idea"*, *"okay"*, *"I agree"*. Similarity, a concept of 'Reject' in Negochat negotiation dialogue corpus [2] can be expressed as *"I disagree"*, *"I reject your proposal"*, *"it's not accepted"*.

Historically, the problem of lexical variability was handled by collecting more training data. While highly effective, the data collection procedure is also unfortunately very expensive and time-consuming to perform. Therefore, we focus on exploiting lexical resources with a typical relatively small training set size.

Paraphrasing shows its importance in many areas of NLP, such as question expansion in question answering [3], summarization [4], text simplification [5], and sentence similarity in machine translation (MT) [6]. Paraphrase expansion (PE) has improved the expressiveness of the natural language generation (NLG) [7] and intent classification in natural language understanding (NLU) [8], [9].

Various approaches have been considered for paraphrase generation. These methods greatly differ in their complexity and the amount of NLP resources that they depends on. They vary from rule-based paraphrase generation methods [10] and thesaurus-based methods [6] to NLG-based and SMT-based methods. In NLG-based methods, the source sentence is encoded in some semantic representation, then the NLG system is employed to generate a natural language sentence from the given semantic representation [11], [12]. SMT-based methods that were used as monolingual MT, translate the source sentences into target sentence that is in the same language [13]. We will generate paraphrases by using the extension of the pivot approach to extracting paraphrase phrases [14]. This method was initially proposed in [3].

In this work, we explore how the automatic paraphrase expansion helps to address lexical variability. In addition, we define the criterion that the "best" pivot language should satisfy. We show the contribution of the PE on the task of intent classification in dialogue systems (DS). The rest of this paper is structured as follows. We review related works on resolving lexical variability via paraphrase expansion in Section II. The formal definition of the multi-pivot approach to generate paraphrases is defined in Section III. Our method that finds effective parameters for paraphrase expansion is described in Section IV. Section V presents the extrinsic evaluation on the intent classification task. Finally, our analysis is in Section VI.

II. Related works

The milestone work that introduced the multi-pivot approach to generate paraphrases via MT was done in [3]. The existing MT systems were used to generate semantically equivalent, but lexically and syntactically distinct, questions. They showed that the contribution of paraphrase expansion to QA performance is significant, and replacing a question with a more lexically rich question can result in a 35% performance increase.

Another interesting work that applied the similar methodology to generate paraphrases showed that enriching SMT training data with carefully selected paraphrases can improve the performance of SMT on both the training corpus of small and medium size [15].

Enriching the training corpus with manually added paraphrases increases the accuracy of the intent classification [8] in DS designed for tactical questioning training [16]. For each of the 296 utterances in the 19 dialogues, they had nine annotators to create a set of paraphrases. The enriched training corpus improved the weak accuracy from 56% to 67%, approaching the level of human performance.

Another interesting work is based on Mission Plastechnologie situated dialogue system [9], those training corpus was enriched with 38000 automatically generated paraphrases. The paraphrases were generated by the multi-pivot machine

ISBN 978-952-68397-8-3 (PDF)

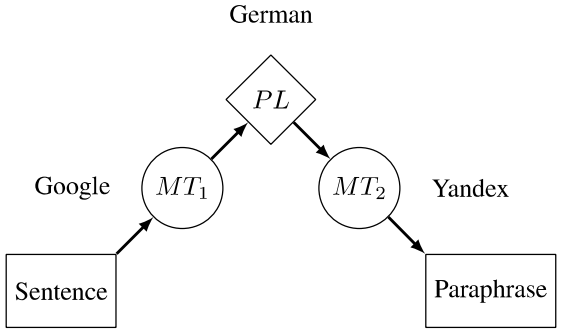

Fig. 1. A single-pivot paraphrase expansion scheme – (Google, German, Yandex)

translation approach with six languages (English, Spanish, Italian, German, Chinese and Arabic), but with only a single MT engine, even though this expansion yielded an increase of up to 8 points in combination with lemmatization.

III. Multi-pivot approach for paraphrase generation

The formal definition of the pivot approach based on parallel multilingual corpora was adopted from [17]. The assumption is that two English phrases that are translated to the same phrase in a foreign language (a pivot language) are potential paraphrases.

A single-pivot PG system is defined a s a triple (MT_1, PL, MT_2), where a MT engine MT_1 translates a source sentence S into a pivot language PL and then MT engine MT_2 translates the resulting sentence back into the source language. Fig. 1 depicts a single-pivot paraphrase generation scheme – (Google, German, Yandex).

A set of single-pivot systems with various pivot languages and MT engines is a multi-pivot PG system. With m pivot languages and n MT engines, we can build a multi-pivot paraphrase generating system consisting of $N = n^2 \times m$ single-pivot ones.

In this work we used nine pivot languages: Portuguese, French, German, Hebrew, Russian, Arabic, Finish, Chinese, Hungarian and three MT engines: Google Translate API, Microsoft Translator Text API, Yandex Translate API. Therefore, we have a multi-pivot PG system consisting of 81 $(3 \times 9 \times 3)$ single-pivot systems.

IV. Performing effective paraphrase expansion

While researching the contribution of the multi-pivot paraphrase expansion, we want to answer the two following questions:

- What is the best performing pivot language among all given MT engines' combinations?
- What is the best performing combination of MT engines among all given pivot languages?

The answer to the first question allows us to pick the best performing pivot language (if there is a limited number of allowed pivot language to use). In the same way, the answer to the second question points out to the best combination of the MT engines. Both answers save the priceless resource to generate the paraphrases, and time to train the classification model.

Before answering the first question empirically, we define the criteria for the "best" pivot language to generate paraphrases:

- MT engines should provide high-quality translations for a pivot language (semantic equivalence criterion).
- A pivot language should introduce lexical variability to the paraphrases (lexical dissimilarity criterion).

The above criteria follow from the standard criteria to measure the paraphrase quality. It was shown that there is no automatic metric that is capable to measure all these criteria in paraphrase generation [18]. Here we state that by careful analysis of the available pivot languages we can choose one that at least for the lexical dissimilarity criterion achieves the higher value.

In this work we do not measure the quality of machine translations, but there are a couple of ways to smooth out the problem of semantic equivalence.

- Use MT engines that are recommended for production environment.
- Choose pivot languages with enough training data.
- On shorter source sentences it is easier to ensure high-quality translations.
- Using only lexical features in the classification model does not require perfect syntactic correctness of the target paraphrase.

In order to measure lexical dissimilarity of the paraphrases we use BLEU measure. BLEU has been the most widely used metric for machine translation evaluation [19]. BLEU's output is a number between 0 and 1. This value indicates how similar the two texts are, with values closer to 1 representing more similar texts. In this work, we used the smoothed version of the BLEU measure, that was initially proposed in [20]. More techniques to calculate the sentence-level BLEU can be found in [21].

To measure how similar a translation T and its reference R, we calculate BLEU by multiplying precision $P(N, T, R)$ and brevity penalty $BP(T, R)$:

$$BLEU(N, T, R) = P(N, T, R) \times BP(T, R) \qquad (1)$$

where $P(N, T, R)$ is the geometric mean of n-gram precisions:

$$P(N, T, R) = \left(\prod_{n=1}^{N} p_n \right)^{\frac{1}{N}}, \text{ where } p_n = \frac{m_n}{l_n} \qquad (2)$$

m_n is the number of matched n-grams between translation T and its reference R, and l_n is the total number of n-grams

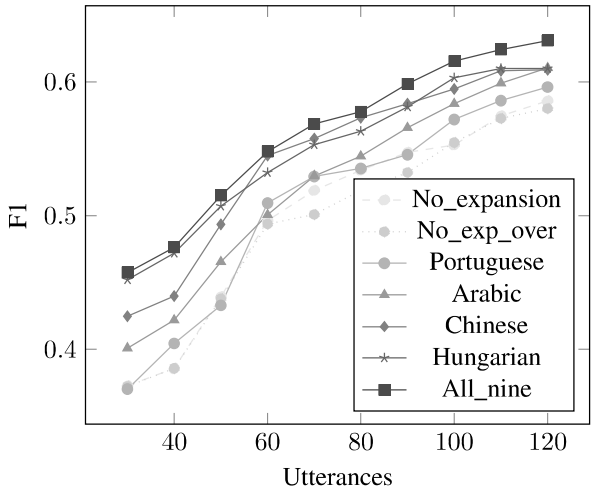

Fig. 2. The macro-averaged F1 performance of intent classification

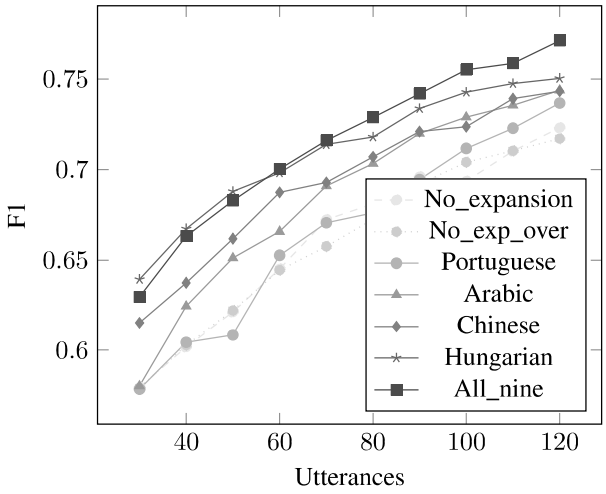

Fig. 3. The micro-averaged F1 performance of intent classification

in the translation T. If the translation length $len(T)$ is shorter than the reference length $len(R)$, then brevity penalty lowers the score.

$$BP(T,R) = min\left(1.0, exp\left(1 - \frac{len(R)}{len(T)}\right)\right) \quad (3)$$

The applied smoothing technique adds 1 to the matched n-gram count and the total n-gram count for n ranging from 2 to N.

$$m'_n = m_n + 1,\ l'_n = l_n + 1 \text{ for } n \text{ in } 2...N \quad (4)$$

Then we calculate the smoothed BLEU measure between each paraphrase and its source utterance. Finally, for each pivot language we compute the average of the BLEU measures of all paraphrases that were generated via the given pivot language (we performed this procedure among all MT combinations). Our assumption is that the pivot language with the highest dissimilarity should outperform the other pivot languages in extrinsic evaluation.

V. Extrinsic evaluation

We evaluated the contribution of the paraphrase expansion on intent classification task. We test our approach on Negochat negotiation corpus [2].

The *job-candidate* domain of the Negochat negotiation corpus was used in several previous works [22], [23], [2], [24]. The negotiation takes place between an *employer* and a *candidate*. In the corpus, the agent is always the *candidate*, while the human is the *employer*. The goal of both sides is to reach a consensus on several issues in order to sign a hiring agreement, while optimizing their own score objective. The issues discussed are *salary*, *working hours*, *car*, *pension*, *promotion* and *position*. The closed list of negotiation issues with their predefined set of values can be found in [2]. The negotiation dialogue comprises the exchange of one or more of the following intent in each utterance: 'Offer', 'Accept', 'Reject', 'Query', 'Greet', 'Quit', where all but the last two are considered crucial for maintaining the dialogue flow. The Negochat corpus comprises 100 dialogues with 1409 human utterances. Each utterance in this corpus is labeled with one or more classes and an associated issue and value. For example, the utterance *"I offer you a pension of 10%"* is labeled as (*class=Offer, issue=Pension fund, value=10%)*. The collected data hardly contain any lexical variability in issues and values, which can be explained by the fact that the set of issues and corresponding values is closed. Therefore, we discarded the parts of the text that describe the issues and their values, in order to let the classifier focus more easily on the words that express intents (for example, the utterance *"I offer you a salary of 90,000"* is converted into *"I offer you"*). The task of composite semantic label classification reduces to intent classification. The distribution of the intents is highly unbalanced: 'Offer' - 59%, 'Reject' - 17%, 'Accept' - 16%, 'Query' - 5%, 'Greet' - 2%, 'Quit' - 1%.

We enriched the Negochat corpus with automatically generated paraphrases (the enriched corpus is freely available at https://github.com/vaskonov/ainl). Then we calculated averaged smoothed BLEU value of all nine pivot languages paraphrases according to the Section IV. The averaged smoothed BLEU measures is in Table I.

TABLE I. The averaged smoothed BLEU measure in sorted order

Hungarian	0.52
Chinese	0.54
Finish	0.57
Arabic	0.59
Hebrew	0.60
Russian	0.62
French	0.67
German	0.68
Portuguese	0.69

We used the same methodology to compare the performance on intent classification as in [24]. We leveraged an SVM classifier (a state of the art classifier for intent classification [25]). The classifier converts the text of an utterance into a set of input features. As input features to the classifier we used unigrams. In this work, we refrained from using the

context features, because we wanted to concentrate on the lexical contribution of the paraphrase expansion.

We compared the performance with a 10-fold cross validation experiment, as follows. The dataset was divided into 10 folds. Each time, a single fold was designated as the training set and the remaining 9 folds as the test set.

The following training datasets were compared:

No expansion The initial Negochat dialogues were used as a training set without paraphrase expansion.

Single pivot language expansion We generated four datasets with one single pivot language expansion of the four following pivot language (Portuguese, Arabic, Hungarian, Chinese). The resulting datasets are nine times bigger than the initial (for every initial utterance we have 9 paraphrases).

No expansion oversampled We oversampled the initial dataset by taking every utterance ten times, such than the size of the oversampled dataset equals the size of the expanded dataset by a single pivot language.

All nine pivot languages expansion The initial Negochat dialogues were expanded by all nine available pivot language paraphrases.

Due to the fact that the distribution is unbalanced we calculated macro-average metrics. Macro-average measures give equal importance to classes of different frequencies. For completeness, we also report performance using the micro-average measure, which gives more weight to frequent classes [26]. We omit the 'Greet' and 'Quit' intents from the calculations, because they are not crucial for the dialogue flow and there is an insufficient number of utterances labeled as 'Greet' and 'Quit'.

VI. Analyses

Table I shows the averaged smoothed BLEU per pivot language. The data show that Hungarian gets the highest lexical dissimilarity and the lowest is Portuguese.

As shown in Fig. 2 the macro-averaged F1 performance correlates well with the averaged smoothed BLEU measure. As expected, the most dissimilar pivot language notably outperforms the other pivot languages. Besides, the performance of the Hungarian language almost coincides with the performance of all nine pivot languages expansion. This means that by expanding the dataset with the most dissimilar pivot language we can save the expenses to generate paraphrases for additional pivot languages and save the priceless resources to train a classification model. The micro-averaged F1 performance in Fig. 3 has a similar trend.

Another interesting point to mention is that the languages from the same language family are located very close to each other in the Table I. For example, the Semitic languages (Arabic and Hebrew) have similar averaged smoothed BLEU measure. The Uralic languages (Finish and Hungarian) is not far from each other too. Two Romance languages (Portuguese and French) are very close to each other. This proximity can be explained by the comparable nature of the languages from the same language family that is reflected in the machine translation model.

In addition, we identified the best performing pivot language per intent. It turns out, that Hungarian is the best performing single pivot language for 'Offer', 'Accept' and 'Reject' and Chinese is the best for 'Query' (more details at https://github.com/vaskonov/ainl).

The same calculation of the averaged smoothed BLEU measure for the nine MT engines combinations shows that the resulting BLEU values are very close to each other. However, the combinations with identical engines, as expected, lead to a slightly higher BLEU measure (the identical engines introduce minimal amount of the lexical variability). The combination of (Yandex, Yandex) has the highest average smoothed BLEU of 0.65, then goes (Google, Google) with 0.64, surprisingly the combination of (Microsoft, Microsoft) has 0.59 and it falls into the range of mixed MT engines with very similar BLEU values that differ from 0.57 to 0.60. The empirical evaluation of the different combinations of the MT engines shows that the difference between their performance is insignificant (more details at https://github.com/vaskonov/ainl). This means that the difference between pivot languages and their corresponding training corpora for MT engines is much more significant than the difference between the translation model of different MT engines.

The same evaluation was performed on the sample of the Switchboard corpus with a similar conclusion (more details at https://github.com/vaskonov/ainl).

Acknowledgements

We thank Sarit Kraus, Oren Melamud, Ron Artstein, Avi Rosenfeld, Erel Segal-Halevi, Osnat Drein and Inon Zuckerman for their helpful comments and remarks. This work was partly supported by ERC Grant #267523.

VII. Conclusion

Our experiments show that the paraphrase expansion leads to better performance than the BoW baseline. In addition, we found that the performance of the different pivot languages correlates well with the averaged smoothed BLEU measure. We defined the desired criteria for the pivot language. In order to ensure the semantic similarity with the source sentence, the paraphrase generator should use the language with enough training data. However, the chosen pivot language should be far enough from the English in order to introduce lexical variability. According to our experiment, it does not matter which MT engine combination is used, however, we advise to use the combination of different MT engines in order to ensure lexical variability. We found that by using the most suitable pivot language we could save the expenses to generate the paraphrases for an addition pivot languages and save the resource to generate a classification model.

In this work, we showed that the languages from the same language family have similar averaged smoothed BLEU value. In future work, we would like to explore the contribution of paraphrase expansion by language families. The optimal combination of the pivot languages potentially could lead to a better performance than the single farthest pivot language from English. In addition, we would like to test the proposed approach on other types of tasks. Finally, it is important to come up with a metric that could concurrently evaluate the

two different criteria for the best pivot language (semantic equivalence and lexical dissimilarity).

To promote further research, we made our code freely available (https://github.com/vaskonov/).

REFERENCES

[1] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[2] V. Konovalov, R. Artstein, O. Melamud, and I. Dagan, "The negochat corpus of human-agent negotiation dialogues," in *LREC*, 2016.

[3] P. A. Duboue and J. Chu-Carroll, "Answering the question you wish they had asked: The impact of paraphrasing for question answering," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, 2006, pp. 33–36.

[4] L. Zhou, C.-Y. Lin, D. S. Munteanu, and E. Hovy, "Paraeval: Using paraphrases to evaluate summaries automatically," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 447–454.

[5] J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait, "Simplifying text for language-impaired readers," in *Proceedings of EACL*, vol. 99, 1999, pp. 269–270.

[6] D. Kauchak and R. Barzilay, "Paraphrasing for automatic evaluation," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 455–462.

[7] L. Iordanskaja, R. Kittredge, and A. Polguere, "Lexical selection and paraphrase in a meaning-text generation model," in *Natural language generation in artificial intelligence and computational linguistics*. Springer, 1991, pp. 293–312.

[8] D. DeVault, A. Leuski, and K. Sagae, "Toward learning and evaluation of dialogue policies with text examples," in *Proceedings of the SIGDIAL 2011 Conference*. Association for Computational Linguistics, 2011, pp. 39–48.

[9] C. Gardent and L. M. R. Barahona, "Using paraphrases and lexical semantics to improve the accuracy and the robustness of supervised models in situated dialogue systems," in *Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 808–813.

[10] D. Lin and P. Pantel, "Discovery of inference rules for question-answering," *Natural Language Engineering*, vol. 7, no. 04, pp. 343–360, 2001.

[11] R. Kozlowski, K. F. McCoy, and K. Vijay-Shanker, "Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources," in *Proceedings of the second international workshop on Paraphrasing-Volume 16*. Association for Computational Linguistics, 2003, pp. 1–8.

[12] R. Power and D. Scott, "Automatic generation of large-scale paraphrases," in *Third International Workshop on Paraphrasing*, 2005, pp. 73–79.

[13] C. Quirk, C. Brockett, and W. B. Dolan, "Monolingual machine translation for paraphrase generation." in *EMNLP*, 2004, pp. 142–149.

[14] C. Bannard and C. Callison-Burch, "Paraphrasing with bilingual parallel corpora," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 597–604.

[15] W. He, S. Zhao, H. Wang, and T. Liu, "Enriching smt training data via paraphrasing." in *IJCNLP*. Citeseer, 2011, pp. 803–810.

[16] D. Traum, A. Leuski, A. Roque, S. Gandhe, D. DeVault, J. Gerten, S. Robinson, and B. Martinovski, "Natural language dialogue architectures for tactical questioning characters," DTIC Document, Tech. Rep., 2008.

[17] S. Zhao, H. Wang, X. Lan, and T. Liu, "Leveraging multiple mt engines for paraphrase generation," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 1326–1334.

[18] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 190–200.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[20] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 605.

[21] B. Chen and C. Cherry, "A systematic comparison of smoothing techniques for sentence-level bleu," *ACL 2014*, p. 362, 2014.

[22] R. Lin, S. Kraus, J. Wilkenfeld, and J. Barry, "Negotiating with bounded rational agents in environments with incomplete information using an automated agent," *Artificial Intelligence*, vol. 172, no. 6, pp. 823–851, 2008.

[23] Y. Oshrat, R. Lin, and S. Kraus, "Facing the challenge of human-agent negotiations via effective general opponent modeling," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 377–384.

[24] V. Konovalov, O. Melamud, R. Artstein, and I. Dagan, "Collecting better training data using biased agent policies in negotiations," in *WOCHAT*, 2016.

[25] G. Tur, D. Hakkani-Tur, and L. Heck, "What is left to be understood in atis?" in *Spoken Language Technology Workshop (SLT), 2010 IEEE*. IEEE, 2010, pp. 19–24.

[26] H. Sch¨utze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.