Semantic Equivalence Detection: Are Interrogatives Harder than Declaratives?

João Rodrigues, Chakaveh Saedi, António Branco and João Silva
University of Lisbon
NLX-Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências
Campo Grande, 1749-016 Lisboa, Portugal
{joao.rodrigues, chakaveh.saedi, antonio.branco, jsilva}@di.fc.ul.pt

Abstract

Duplicate Question Detection (DQD) is a Natural Language Processing task under active research, with applications to fields like Community Question Answering and Information Retrieval. While DQD falls under the umbrella of Semantic Text Similarity (STS), these are often not seen as similar tasks of semantic equivalence detection, with STS being implicitly understood as concerning only declarative sentences. Nevertheless, approaches to STS have been applied to DQD and paraphrase detection, that is to interrogatives and declaratives, alike. We present a study that seeks to assess, under conditions of comparability, the possible different performance of state-of-the-art approaches to STS over different types of textual segments, including most notably declaratives and interrogatives. This paper contributes to a better understanding of current mainstream methods for semantic equivalence detection, and to a better appreciation of the different results reported in the literature when these are obtained from different data sets with different types of textual segments. Importantly, it contributes also with results concerning how data sets containing textual segments of a certain type can be used to leverage the performance of resolvers for segments of other types.

Keywords: semantic text similarity, paraphrase detection, duplicate question detection

1. Introduction

Semantic Text Similarity (STS) is a Natural Language Processing (NLP) task whereby a system, given two input text segments, assigns to them a similarity score in a discrete or continuous scale that ranges from representing total similarity—for semantically equivalent segments—to representing total dissimilarity—for segments that are semantically independent.

The STS task has been part of the SemEval competitive shared tasks since 2012 (Agirre et al., 2012), together with other challenges for a wide variety of other tasks, such as plagiarism detection, sentiment analysis or relation extraction, to name but a few. More recently, SemEval embraced STS challenges that concern more focused tasks, like paraphrase detection, which consists of a binary decision on whether two input sentences are paraphrases of each other and, starting in 2016, a task on Duplicate Question Detection (DQD) (Nakov et al., 2016).

DQD appears as a special case of paraphrase detection, where the focus is on interrogative sentences: this task consists of a binary decision on whether two input interrogatives sentences are a duplicate of each other.

The motivation for the increasing interest in DQD, and the inclusion in SemEval of challenges dedicated to DQD, comes from the increasing popularity of on-line Community Question Answering (CQA) forums, such as Stack Exchange[1] or Quora[2]. These forums are quite open in allowing any user to post questions (and answer questions from other users) but from this arises a potential problem that may eventually affect the effectiveness of these on-line services, namely that many posted questions are duplicates of questions already answered. In such cases, the user posting the duplicate question should be directed to the already existing question. Duplicate questions are manually flagged by the users, but this effort quickly becomes unwieldy as the site grows in popularity, driving the need for automatic procedures for DQD.

Though the interest in DQD may be seen as relatively recent, there is an accumulated body of lessons learned about this task and the expected performance of systems tackling it, some of them being quite in line with what is known about data-driven approaches in general, while some others are more specific for this task. From existing work on DQD, such as (Bogdanova et al., 2015) (Rodrigues et al., 2017) and (Saedi et al., 2017), one learned that (i) training and evaluating over a specific domain with less data, rather than over a generic one with more data, will likely lead to better performance; (ii) training on as much data as possible, gathered from all different domains, and evaluating on a specific domain yields little more than random choice performance; (iii) when training on data sets of interrogative sentences, differences in the average length or in the level of grammaticality of sentences have little impact on performance; (iv) the differences in performance between the major types of approaches to DQD become smaller as the domain becomes more generic; and (v) the best variants of these major approaches all deliver competitive results when trained with general domain data sets with 30,000 sentence pairs, with accuracy scores falling within a range of just 2 to 3 percentage points.

The underlying semantic relation between sentences that STS is seeking to model is the one of synonymy. Interestingly, while it concerns the ultimate notion of semantic equivalence for both types of sentences, declaratives and interrogatives alike, the synonymy relation has quite different operational definitions for each one of them. Two declarative sentences are synonymous (or paraphrases of

[1] http://stackexchange.com/
[2] http://quora.com/

each other) just in case they can replace each other and the truth conditions of any text they happen to be part of are preserved under that substitution (modulo so-called opaque contexts). Two interrogatives, in turn, are synonymous (or duplicates of each other) just in case any successful answer to any one of them is also a successful answer to the other one.

- Declarative duplicate pair (from MSRPC)
  - Dogs, he said, are second only to humans in the thoroughness of medical understanding and research.
  - He said that dogs are second only to humans in terms of being the subject of medical research.
- Interrogative duplicate pair (from DupStack)
  - Where did the notion of "one return only" come from?
  - Should I return from a function early or use an if statement?

Figure 1: Examples of duplicate pairs

The examples in Figure 1 are instances that comply with these two operational definitions of semantic equivalence, for declaratives and for interrogatives. It does not go unnoticed that the superficial similarity—in terms of common words, word order, length, etc.—between the interrogative sentences is much more rarefied than between the declarative ones. And it is on the basis of superficial features that decisions on the eventual underlying relations of semantic equivalence are made. The contrast between these two illustrative pairs of examples thus strongly suggests that, when it comes to STS, we may be facing two tasks of synonymy detection of quite different levels of difficulty, depending on whether we are modeling synonymy between declaratives, or between interrogatives.

Against this background, what has not been studied yet, and remains an interesting research question, is whether the operational and qualitative difference between the synonymy relations for declarative and for interrogative sentences leads to an impact and a substantive difference between the performance of STS systems for declaratives, on the one hand, and for interrogatives, on the other hand. Or, in other words, in what concerns the automatic detection of semantic similarity, are interrogative sentences more difficult to handle than declarative ones given the current methods at hand to tackle them?

This is the driving research question we seek to address and that motivates the experiments reported in the present paper, as well as other subsidiary research questions.

The remainder of this paper is organized as follows. Section 2. describes the related work. We present the data sets and the experimental approaches to undertake STS in Section 3.. In Sections 4., 5. and 6., the experiments carried out are reported and the results of their evaluation are presented, that address the research questions, respectively, whether interrogatives are harder than declaratives, whether merging data sets for different type of textual segments improve the performance of the resolvers, and how difficult are Tweets for semantic equivalence detection. Section 7. concludes with final remarks.

2. Related Work

In a recent study, (Saedi et al., 2017) undertook a systematic comparison of the performance of different major approaches to DQD over progressively larger data sets, by considering approaches that have been identified in the literature as very competitive, namely rule-based, support vector machines classifiers, and deep convolutional neural networks. In the eventual full paper, of which the present document is just an extended abstract, we will present at length these related works.

For now, we restrict ourselves to the key results of that systematic study of the learning curves of these major approaches. A major finding is that there is no approach that beats all others in every point of the learning curve. Simpler, rule-based approaches, like the Jaccard index, are highly competitive for small data sets, but as more data becomes available, they lose out to more sophisticated approaches. In particular, and confirming a widespread assumption in Machine Learning, deep learning approaches come into their own, and its performance surpasses all other approaches, only when a sufficiently large amount of training data is available, containing above 30,000 pairs of sentences with a 50/50 split between duplicates and non-duplicates.

3. Data Sets and Approaches to Semantic Equivalence Detection

This Section describes the data sets and the different approaches to STS that were used in our experiments.

3.1. Data sets

To carry out the proposed experiment, at least two corpora are required, one with interrogative sentences and another with declarative sentences. The corpora should be as close as possible to each other in other aspects, particularly in terms of domain, size and class distribution, in order to obtain results that can be as comparable as possible. To support the testing of our central hypothesis that interrogatives are harder than declaratives in terms of semantic equivalence detection, we resorted to two data sets, one with interrogatives (Quora) and another with declaratives (MSRPC). We also included a third data set, with a mixture of declarative and interrogative sentences (DupStack), and a fourth one (PIT), with segments of highly compromised grammaticality.

These data sets, all in English, are introduced below.

MSRPC is the Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005). It consists of 5,801 pairs of edited and grammatically well-formed declarative sentences taken from news articles on various topics, with each pair being annotated with a binary label indicating if its sentences are paraphrases of each other or not. There are 3900 pairs with paraphrases, and the sentences are on average 18.92 words long.

Quora is a corpus originating from the CQA forum Quora (Iyer et al., 2017). It contains over 404,289 pairs of edited and grammatically well-formed questions, annotated as to whether they contain duplicates or non-duplicates, of which 149,263 are pairs with duplicates. These questions address any topic and on average are 11.06 words long.

DupStack is the acronym used in this paper for CQADupStack (Hoogeveen et al., 2015), a corpus composed of pairs of threads from StackExchange, another popular CQA site, annotated with information as to whether they are a duplicate. The 3,891,016,008 pairs of threads are sourced from 12 subforums of StackExchange resorting to the original classification split, covering a number of subtopics mostly in the ICT domain, of which 10,677 contain duplicate pairs. These segments have an average length of 8.48 words, with some of them presenting sub-optimal grammaticality. They were entered by the users as the title of a larger text where the information being searched for is indicated, and thus with many of them appearing clearly as declaratives.

PIT is the corpus Paraphrase and Semantic Similarity in Twitter for the SemEval-2015 Task 1 (Xu et al., 2015). It contains over 18,000 pairs of segments, taken from Twitter. These sentences tend thus to be relatively short and make extensive use of abbreviations and a highly compromised grammaticality. There are 5,641 pairs with paraphrases, and the segments have on average a length of 8.13 words.

For the sake of comparability of the experimental results to be obtained on the basis of these data sets, we take an equal number of sentence pairs from each corpus, randomly selected, but ensuring a balanced distribution with an equal number of duplicate and non-duplicate cases. The smallest corpus, MSRPC, with 3,900 duplicate (and 1,901 non-duplicate) pairs, constraints the maximum number of sentence pairs that can be picked. Accordingly, from each corpus, 3,900 duplicate and 3,900 non-duplicate pairs are randomly selected, for a total size of 7,800 pairs per corpus. Note that MSRPC only has 1,901 non-duplicate pairs. The other 1,999 non-duplicate pairs in the respective sub-corpus are generated by randomly pairing sentences taken from distinct pairs.

In all experiments, 80% of the pairs are used for training and 20% for testing.

Table 1 summarizes information on the type and size of the sub-corpora used.

	type	grammaticality	#tokens
MSRPC	declar.	ok	301,428
Quora	interrog.	ok	177,334
DupStack	mixed	sub-optimal	142,387
PIT	tweets	highly compromised	137,898

Table 1: The four sub-corpora, each with 7,800 pairs.

While the number of sentence pairs is the same for all corpora, MSRPC has a much higher number of tokens. This happens because the sentences in that corpus, which are taken from news articles, are usually longer than the questions from Quora and DupStack, or the tweets from the PIT corpus.

3.2. Approaches to semantic equivalence detection

We use the same set of approaches for DQD from (Rodrigues et al., 2017) and (Saedi et al., 2017), as these cover a range of different methods with state-of-the-art performance for the size of training data there are available for the present experiments. Given the hypothesis that detecting synonymy between interrogatives is harder than between declaratives, we resort to approaches to semantic equivalence detection with highly competitive performance for DQD, in order to explain away a possible justification for the difference in performance between the two types of sentences based on the putative weakness of the methods used vis a vis interrogatives.

In this extended abstract we provide a short summary of each approach—which will be extended in the eventual full paper—, and for now, direct the reader to the articles cited above for further information.

Jaccard The Jaccard index is a straightforward statistic based on the count of common of n-grams between the two segments being compared. It is used as a simple baseline that previous work has shown to nonetheless be very competitive (Wu et al., 2011), especially for small sized data sets below 30,000 pairs (Saedi et al., 2017). All n-grams, with n ranging from 1 to 4, are used.

SVM Support Vector Machine classifiers have been used with success in many NLP tasks and are able to cope with a great variety of features. The set of features used in this work is formed by (i) two vectors with the one-hot encodings of n-gram occurrences in each segment; (ii) the Jaccard index scores for 1, 2, 3 and 4-grams; (iii) the counts of negative words (e.g. *never*, *nothing*, etc.) in each segment; (iv) the number of nouns that are common to both segments; and (v) the cosine similarity between the vector representation of each segment.

DCNN Deep Neural Networks have, over the past few years, gained popularity and been applied to many NLP tasks, often surpassing by a large margin the other alternative approaches if sufficiently large training data is available. In this work, we use the architecture introduced in (Rodrigues et al., 2017), which combines a convolutional neural network and a deep network in a Siamese architecture.

4. Are Interrogatives harder than Declaratives?

In this Section, we present data and experiments whose results are suited to bring empirical evidence that can support our research hypothesis that interrogatives are harder than declaratives in terms of semantic equivalence detection.

4.1. Semantic equivalence

A first and straightforward experiment consists of training and evaluating each one of the three working approaches for semantic equivalence detection over each one of the two principal data sets, the ones with declaratives (MSRPC) and with interrogatives (Quora). The results are summarized in Table 2. Note that for SVM and DCNN, the scores shown are the average of 3 runs.

	Jaccard	SVM	DCNN
Declaratives (MSRPC)	**77.30**	**80.46**	**78.42**
Interrogatives (Quora)	69.29	69.10	72.78

Table 2: Accuracy scores (%) of the different approaches applied to the two different types of segments, with the **highest values** (in bold) for declaratives.

The three approaches perform better for declaratives than for interrogatives across the board by a substantial margin, that ranges from over 5 (DCNN) to over 11 percentage points (SVM). This provides clear empirical support for our research hypothesis.

4.2. Superficial overlap

As noted in Section 1., a major motivation to put forward our research question is the observation of the different operational definitions for the semantic equivalence of declaratives and of interrogatives. Under these definitions, for two sentences to be equivalent, less superficial commonalities between them seem to be expected to hold on average for interrogatives than for declaratives.
A way to obtain a quantitative test for this expectation is to measure the level of superficial overlap. For this purpose, we calculate the average BLEU score between the sentences in equivalent pairs and between the sentences in non-equivalent pairs, where higher scores reflect higher overlap. The results are reported in Table 3.

	equivalent	non-equivalent
Declaratives (MSRPC)	49.73	18.93
Interrogatives (Quora)	31.20	17.33

Table 3: Averaged BLEU scores of sentences in equivalent and non-equivalent pairs

With the highest score for declaratives (49.73), and a difference to interrogatives of over 18 BLEU points, for sentences that are semantic equivalents of each other, these scores provide an objective confirmation that equivalent interrogatives have less superficial overlap than equivalent declaratives. Hence, this grants objective support for the expectation that interrogatives should be harder than declaratives for semantic equivalence detection, as the mainstream approaches for this task are data-driven and rely on the superficial similarity of sentences for their operation.

4.3. Mix of types

An additional piece of evidence that may support our research hypothesis can be looked for in the performance of the equivalence detection systems when running over a data set like DupStack, composed by a mixture of declarative pairs and interrogative pairs, and even mixed-type pairs.

- Duplicate mixed-type pair (from DupStack)
  - Turn-by-turn direction using PgRouting
  - How to emulate Google Maps driving directions using pgRouting?
- Non-duplicate mixed-type pair (from DupStack)
  - SLD: OGC Filter set, but symbolizer expected
  - How to delete coordinate system from raster file with prj.adf?

Figure 2: Example of mixed-type pairs

If our hypothesis holds, the accuracy scores for this data set should lie between the higher and lower scores of declaratives and interrogatives, respectively. The results of this experiment are reported in Table 4.

	Jaccard	SVM	DCNN
Declaratives (MSRPC)	77.30	80.46	78.42
Mixed (DupStack)	**74.16**	**71.23**	81.51
Interrogatives (Quora)	69.29	69.10	**72.78**

Table 4: Accuracy scores (%) of the different approaches applied to the different types of segments, with the **in-between values** (in bold) for the data sets with a mix of declaratives and interrogatives with two of the three approaches.

In two (Jaccard and SVM) of the three approaches, the scores are in line with this prediction, with in-between values. Overall, this provides yet another piece of empirical evidence to the research hypothesis, with the value for DCNN (81.51) appearing as an outlier.

5. Interrogatives and Declaratives Leveraging each other?

The results in Section 4. help to clarify that interrogatives are harder than declaratives in terms of the mainstream approaches to resolve semantic equivalence detection.
Interestingly, this relative advantage gets reverted when it comes to data sets. It is easier to find, collect and support interrogatives with larger data sets, than declaratives. The reason is that semantic equivalent interrogatives happen to be generated in real usage scenarios, as e.g. in the Quora service, making them easy to crowdsource, while declaratives are not.[3] Hence a next research question is to empirically determine whether, and to what extent, the more abundant data sets with interrogatives can help improve the detection of equivalent declaratives.

[3] In the MSRPC corpus, the sentences in the pairs had to be annotated by humans for the specific purpose of the construction of this corpus.

5.1. Cross training

A first experiment seeks to assess how performance is impacted by differences between the training data and the testing data. For this purpose, the overall best system, DCNN, is trained on each corpus and each resulting model is evaluated on each corpus. The results are summarized in Table 5.

	Evaluate on...		
Train on...	MSRPC	DupStack	Quora
Declar. (MSRPC)	**78.42**	56.28	61.67
Mixed (DupStack)	70.58	**81.51**	56.73
Interrog. (Quora)	**76.35**	54.23	72.78

Table 5: Accuracy scores (%) with rows showing training data sets and columns showing evaluation data sets, concerning DCNN

When trained on data sets with declaratives (MSRPC) and with mixed types (DupStack), the best performance is observed when the systems resolve the equivalence for similar types of segments, achieving 78.42 (first row) and 81.51 (second row), respectively.

Very interestingly, when the resolver is trained with the interrogatives (third row), it has the best performance when deciding about declaratives (76.35), and the second best about interrogatives themselves (72.78).

The opposite, however, does not hold. Resolving interrogatives (third column) with systems trained on another type of segments only delivers results that are clearly worse (61.67 and 56.73) than when they are trained on interrogatives themselves (72.78).

This result allows good hopes that the more abundant pairs of duplicate interrogatives can be of help to leverage the performance of resolvers of semantic equivalence between declaratives. Very large data sets with interrogatives collected from real usage scenarios may eventually support the development of resolvers with the best performance than the ones trained on the smaller, and hard to obtain and to expand by explicit manual annotation, data sets containing only declaratives paraphrases. This is the motivation for our next experiment.

5.2. Merged data sets

A second experiment consists thus in training resolvers for declaratives over data sets that contain larger data sets than the MSRPC corpus (with declaratives) alone. These data sets are obtained by resorting to a larger subset of Quora with interrogatives (from 7,800 to 100,000 pairs), and to the merging of this and the other data sets (with 7,800 pairs each) used in this paper.

The results of this experiment are reported in Table 6.

This experiment permits to understand that by growing the size of the training data set with interrogatives by one order of magnitude—from 7,800 to 100,000 pairs, of the new sub corpus Quora100K—is not enough to obtain better results for declaratives (74.42, last row) than when the training data set is smaller—7,800 pairs, of the MSRPC corpus — but made only of declaratives (78.42, penultimate row).

Very interestingly, this experiment allows also to understand that when the larger training data set is obtained by including also the MSRPC data set with declaratives—79.36 by adding Quora100k to it, and 79.42 by adding this and all other data sets to it—, that is enough to overcome the performance of the system trained only with declaratives (78.42, penultimate row).

Train on...	Evaluate on... Decl(MSRPC)
all[1]	79.42
Decl(MSRPC) + Interr(Quora100k)	79.36
Decl(MSRPC)	78.42
Interr(Quora100k)	74.42

[1]Dec(MSRPC) + Int(Quora100k) + Mix(DupStack) + Tw(PIT)

Table 6: Accuracy scores (%) with rows showing training data sets and column showing the evaluation data set, concerning DCNN

These results clearly indicate that adding pairs of interrogatives to the training data set of declaratives is an effective way to improve the performance of paraphrase resolvers. Importantly, this is also a procedure that dispense with a further specific human effort for the construction of the training data set as pairs of duplicate interrogatives can be collected as a byproduct of on-line Community Question Answering forums.

6. How difficult are Tweets after all?

In the context of the results reported in the previous Sections, a third interesting research question to address is to determine how difficult may be the task of semantic equivalence detection for Tweets.

Previous results (Rodrigues et al., 2017) indicate that to a certain extent, reducing the average size of interrogative segments and relaxing the grammaticality of interrogative segments have little impact in equivalence resolvers for interrogatives. But when compared with Tweets—much shorter and with much more compromised grammaticality—these appear as mild differences and the respective may be of little guidance when we turn to Tweets.

- Duplicate pair (from PIT)
  - That 3pointer from Kevin Durant was lucky asf
  - the NBA gods showed favor to that 3
- Non-duplicate pair (from PIT)
  - Aye mac miller new music is aite
  - I swear my waiter is Mac Miller

Figure 3: Examples of Twitter pairs

To answer this third research question, we trained and evaluated the three resolving approaches for semantic equivalence over PIT, the data set with Tweets. Their performance results are reported in Table 7

The three approaches perform worse for Tweets than for the second worst type of segments, viz. interrogatives, by a margin that ranges from almost 1.5 (SVM) to over 20

	Jaccard	SVM	DCNN
Declaratives (MSRPC)	77.30	80.46	78.42
Mixed (DupStack)	74.16	71.23	81.51
Interrogatives (Quora)	69.29	69.10	72.78
Tweets (PIT)	**67.82**	**67.83**	**51.70**

Table 7: Accuracy scores (%) of the different approaches applied to the different types of segments, with the **lowest values** (in bold) for Tweets

percentage points (DCNN). When the resolvers for Tweets are compared to the best resolvers, for grammatical, manually edited declaratives (MSRPC), this gap widens up for a range from about 10 (Jaccard) to almost 30 percentage points (DCNN). This indicates that Tweets are the hardest type of segments in terms of semantic equivalence detection.

7. Conclusions

In this paper we addressed three major research questions related to the task of semantic equivalence detection, and performed a number of experiments that permitted to gather empirical evidence aimed at finding answers for them.

The major driving question is whether interrogatives are harder than declaratives for semantic equivalence resolvers. The higher superficial overlap between declaratives in paraphrasing pairs, than between interrogatives in duplicate pairs, as measured with BLEU; and the substantially superior performance over declaratives, than over interrogatives, of different resolvers developed under major mainstream approaches: all these are major pieces of evidence supporting the observation that this task is harder with interrogatives.

A second research question was whether the performance of resolvers for a given type of segments, interrogative or declarative, can be improved by obtaining larger training data sets that result from the merging of smaller data sets for different types. The contrasting levels of performance of a number of systems developed under these circumstances permitted to observe that this is actually the case with resolvers for declaratives (trained with the merging of data sets with declaratives and interrogatives), but not for interrogatives.

In the context of the previous questions and respective answers, a last research question was the inquiry on how difficult is the task of semantic equivalence detection for Tweets, in comparison to the similar task for grammatically well-formed declaratives and interrogatives. The inferior performance results across the board with different resolvers developed under major mainstream approaches as used in previous experiments permitted to gather empirical evidence indicating that Tweets are the most difficult type of segments for the task of semantic equivalence detection. These results contribute to a better understanding of the strengths and weaknesses of current mainstream methods for semantic equivalence detection, and allow to better appreciate and ponder on the different relevance of the results and scores reported in the literature when these are obtained from different data sets and with different types of textual segments.

8. Acknowledgments

The present research was partly supported by the Infrastructure for the Science and Technology of the Portuguese Language (CLARIN Língua Portuguesa), by the National Infrastructure for Distributed Computing (INCD) of Portugal, and by the ANI/3279/2016 grant.

9. References

Agirre, E., Gonzalez-Agirre, A., Cer, D., and Diab, M. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 385–393.

Bogdanova, D., dos Santos, C. N., Barbosa, L., and Zadrozny, B. (2015). Detecting semantically equivalent questions in online user forums. In *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL)*, pages 123–131.

Nakov, P., Marquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., and Randeree, B. (2016). Semeval-2016 task 3: Community question answering. In *Proceedings of the 11th International Conference on Semantic Evaluation (SemEval)*, pages 27–48.

Rodrigues, J., Saedi, C., Maraev, V., Silva, J., and Branco, A. (2017). Ways of asking and replying in duplicate question detection. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*.

Saedi, C., Rodrigues, J., Silva, J., Branco, A., and Maraev, V. (2017). Learning profiles in duplicate question detection. In *Proceedings of the IEEE 16th International Conference on Information Reuse and Integration (IEEE IRI 2017)*.

Wu, Y., Zhang, Q., and Huang, X. (2011). Efficient near-duplicate detection for Q&A forum. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1001–1009.

10. Language Resource References

Dolan, B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005)*.

Hoogeveen, D., Verspoor, K. M., and Baldwin, T. (2015). CQADupStack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium*, pages 3:1–3:8.

Iyer, S., Dandekar, N., and Csernai, K. (2017). First Quora dataset release: Question pairs.

Xu, W., Callison-Burch, C., and Dolan, B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 1–11.