SAARSHEFF at SemEval-2016 Task 1: Semantic Textual Similarity with Machine Translation Evaluation Metrics and (eXtreme) Boosted Tree Ensembles

Liling Tan[1], **Carolina Scarton**[2], **Lucia Specia**[2] **and Josef van Genabith**[1,3]
Universität des Saarlandes[1] / Campus A2.2, Saarbrücken, Germany
University of Sheffield[2] / Regent Court, 211 Portobello, Sheffield, UK
Deutsches Forschungszentrum für Künstliche Intelligenz[3] / Saarbrücken, Germany
alvations@gmail.com, c.scarton@sheffield.ac.uk,
l.specia@sheffield.ac.uk, josef.van_genabith@dfki.de

Abstract

This paper describes the SAARSHEFF systems that participated in the English Semantic Textual Similarity (STS) task in SemEval-2016. We extend the work on using machine translation (MT) metrics in the STS task by automatically annotating the STS datasets with a variety of MT scores for each pair of text snippets in the STS datasets. We trained our systems using boosted tree ensembles and achieved competitive results that outperforms he median Pearson correlation scores from all participating systems.

1 Introduction

Semantic Textual Similarity (STS) is the task of measuring the degree to which two texts have the same meaning (Agirre et al., 2014). For instance, given the two texts, "*the man is slicing the tape from the box.*" and "*a man is cutting open a box.*", an STS system predicts a real number similarity score on a scale of 0 (no relation) to 5 (semantic equivalence).

This paper presents a collaborative submission between Saarland University and University of Sheffield to the STS English shared task at SemEval-2016. We have submitted three supervised models that predict the similarity scores for the STS task using Machine Translation (MT) evaluation metrics as regression features.

2 Related Work

Previous approaches have applied MT evaluation metrics for the STS task with progressively improving results (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015).

At the pilot English STS-2012 task, Rios et al. (2012) trained a Support Vector Regressor using the lexical overlaps between the surface strings, named entities and semantic role labels and the BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010) scores between the text snippets and their best system scored a Pearson correlation mean of 0.3825. The system underperformed compared to the organizers' baseline system[1] which scored 0.4356.

For the English STS-2013 task, Barrón-Cedeño et al. (2013) also used a Support Vector Regressor with an larger array of machine translation metrics (BLEU, METEOR, ROUGE (Lin and Och, 2004), NIST (Doddington, 2002), TER (Snover et al., 2006)) with measures that compute similarities of dependency and constituency parses (Liu and Gildea, 2005) and semantic roles, discourse representation and explicit semantic analysis (Gabrilovich and Markovitch, 2007) annotations of the text snippets. These similarity measures are packaged in the Asiya toolkit (Giménez and Màrquez, 2010). They scored 0.4037 mean score and performed better than the Takelab baseline (Šarić et al., 2012) at 0.3639.

At the SemEval-2014 Cross-level Semantic Similarity task (Jurgens et al., 2014; Jurgens et al., 2015), participating teams submitted similarity scores for text of different granularity. Huang and Chang (2014) used a linear regressor solely with MT evalu-

[1]Refers to the token cosine baseline system (`baseline-tokencos`) in STS-2012.

Proceedings of SemEval-2016, pages 640–645,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

ation metrics (BLEU, METEOR, ROUGE) to compute the similarity scores between paragraphs and sentences. They scored 0.792 beating the lowest common substring baseline which scored 0.613.

In the SemEval-2015 English STS and Twitter similarity tasks, Bertero and Fung (2015) trained a neural network classifier using (i) lexical similarity features based on WordNet (Miller, 1995), (ii) neural auto-encoders (Socher et al., 2011), syntactic features based on parse tree edit distance (Zhang and Shasha, 1989; Wan et al., 2006) and (iii) MT evaluation metrics, viz. BLEU, TER, SEPIA (Habash and Elkholy, 2008), BADGER (Parker, 2008) and MEANT (Lo et al., 2012).

For the classic English STS task in SemEval-2015, Tan et al. (2015) used a range of MT evaluation metrics based on lexical (surface n-gram overlaps), syntactic (shallow parsing similarity) and semantic features (METEOR variants) to train a Bayesian ridge regressor. Their best system achieved 0.7275 mean Pearson correlation outperforming the `token-cos` baseline which scored 0.5871 while the top system (Sultan et al., 2015) achieved 0.8015.

Another notable mention of MT technology in the STS tasks is the use of referential translation machines to predict and derive features instead of using MT evaluation metrics (Biçici and van Genabith, 2013; Biçici and Way, 2014; Bicici, 2015).

3 Approach

Following the success of systems that use MT evaluation metrics, we train three regression models using an array of MT metrics based on lexical, syntactic and semantic features.

3.1 Feature Matrix

Machine translation evaluation metrics utilize various degrees of lexical, syntactic and semantic information. Each metric considers several features that compute the translation quality by comparing a translation against one or several reference translations.

We trained our system using the follow feature sets: (i) n-gram, shallow parsing and named entity overlaps (`Asiya`), (ii) `BEER`, (iii) `METEOR` and (iv) `ReVal`.

3.1.1 `Asiya` Features

Gonzàlez et al. (2014) introduced a range of language independent metrics relying on n-gram overlaps similar to the modified $n-$-gram precisions of the BLEU metric (Papineni et al., 2002). Different from BLEU, Gonzàlez et al. (2014) computes n-gram overlaps using similarity coefficients instead of proportions. We use the `Asiya` toolkit (Giménez and Màrquez, 2010) to annotate the dataset with the similarity coefficients of n-gram overlap features described in this section.

We use 16 features from both cosine similarity and Jaccard Index coefficients of the character-level and token-level n-grams from the order of bigrams to 5-grams. Additionally, we use the Jaccard similarity of the pseudo-cognates and the ratio of n-gram length as the 17th and 18th features.

Adding a syntactic dimension to our feature set, we use 52 shallow parsing features described in (Tan et al., 2015); they measure the similarity coefficients from the n-gram overlaps of the lexicalized shallow parsing (aka chunking) annotations. As for semantics, we use 44 similarity coefficients from Named Entity (NE) annotation overlaps between two texts.

After some feature analysis, we found that 22 out of the 44 NE n-gram overlap features and 1 of the shallow parsing features have extremely low variance across all sentence pairs in the training data. We removed these features before training our models.

3.1.2 `BEER` Features

Stanojevic and Simaan (2014) presents an MT evaluation metric that uses character n-gram overlaps, the Kendall tau distance of the monotonic word order (Isozaki et al., 2010; Birch and Osborne, 2010) and abstract ordering patterns from tree factorization of permutations (Zhang and Gildea, 2007).

While Asiya features are agnostic to word classes, BEER differentiates between function words and non-function words when calculating its adequacy features.

3.1.3 `METEOR` Features

METEOR first aligns the translation to its reference, then it uses the unigram mapping to see whether they match based on their surface forms,

	answer-answer	headlines	plagiarism	postediting	question-question	All
`Linear`	0.31539	0.76551	0.82063	0.83329	0.73987	0.68923
`Boosted`	0.37717	0.77183	0.81529	0.84528	0.66825	0.69259
`XGBoost`	**0.47716**	**0.78848**	**0.83212**	**0.84960**	**0.69815**	**0.72693**
Median	0.48018	0.76439	0.78949	0.81241	0.57140	0.68923
Best	0.69235	0.82749	0.84138	0.86690	0.74705	0.77807

Table 1: Pearson Correlation Results for English STS Task at SemEval-2016

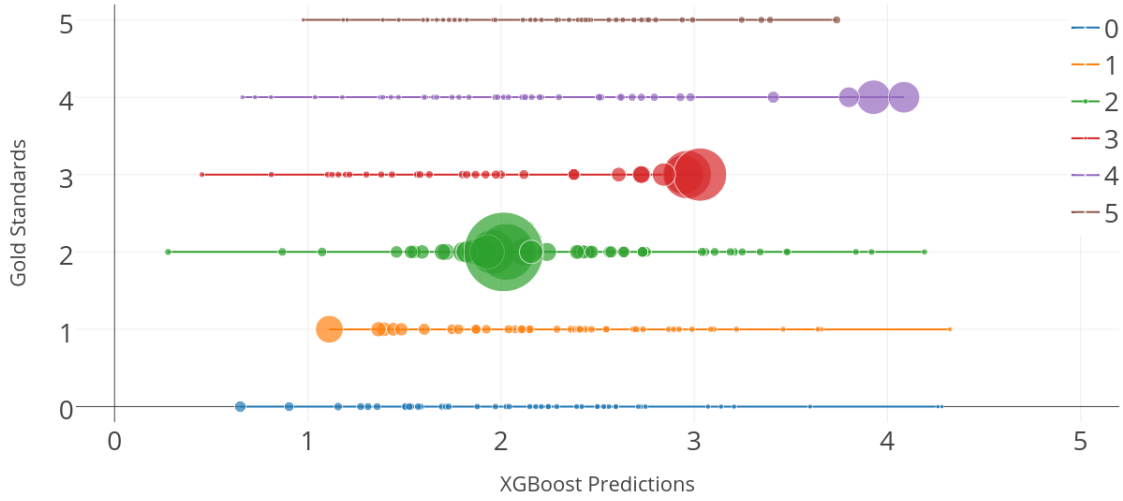

Figure 1: L1 Error Analysis on the answer-answer domain

word stems, synonyms and paraphrases (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010).

Similar to BEER features, METEOR makes a distinction between content words and function words and its recall mechanism weights them differently. We use all four variants of METEOR: exact, stem, synonym and paraphrase.

3.1.4 `ReVal` Features

ReVal (Gupta et al., 2015) is a deep neural net based metric which uses the cosine similarity score between the Tree-based Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997; Cho et al., 2014; Tai et al., 2015) dense vector space representations of two sentences.

3.2 Models

We annotated the STS 2012 to 2015 datasets with the features as described in Section 3.1 and submitted three models to the SemEval-2016 English STS Task using (i) a linear regressor (`Linear`), (ii) boosted tree regressor (`Boosted`) (Friedman, 2001) and (iii) eXtreme Gradient Boosted tree regressor (`XGBoost`) (Chen and He, 2015; Chen and Guestrin, 2015). They were trained using all features described in Section 3.

We have released the MT metrics annotations of the STS data and implementation of systems on `https://github.com/alvations/stasis/blob/master/notebooks/ARMOR.ipynb`

4 Results

Table 1 presents the official results for our submissions to the English STS task. The bottom part of the table presents the median and the best correlation results across all participating teams for the respective domains.

Our baseline linear model outperforms the median scores for all domains except the *answer-answer* domain. Our boosted tree model performs better than the linear model and the extreme gradient boosted tree model performs the best of the three. We note that our correlation scores for all three models is lower than the median for the *answer-answer* domain.

Figure 1 shows the bubble chart of the L1 error analysis of our XGBoost model against the gold standard similarity scores for the answer-answer domain. The colored lines correspond to the integer annotations, e.g. the yellow line represents the data points where the gold-standard annotations are 1.0. The span of the line represents the span of predictions our model made for these texts. The size of the bubble represents the effect size of our predictions' contribution to the Pearson correlation score, i.e. how close our predictions are to the gold standards.

5 Discussion

As we see from Figure 1, the centroids of the bubbles represents our model's best predictions. Our predictions for texts that are annotated at 1 to 4 similarity scores are reasonably close to the gold standards but the model performs poorly for texts annotated with the 0 and 5 similarity scores.

Looking at the texts that are rated 0, we see that there are cases where the n-grams within these texts are lexically / syntactically similar but the meaning of the texts are disparate. For example, this pair of text snippets, *'You don't have to know'* and *'You don't have equipments/facilities'* are rated 0 in the gold standards but from a machine translation perspective, a translator would have to do little work to change *'to know'* to *'equipments/facilities'*.

Because of this, machine translation metrics would rate the texts as being similar and even suitable for post-editing. However, the STS task focuses only on the meaning of the text which corresponds more to the adequacy aspect of the machine translation metrics. Semantic adequacy is often overlooked in machine translation because our mass reliance on BLEU scores to measure the goodness of translation with little considerations for penalizing semantic divergence between the translation and its reference.

On the other end of the spectrum, machine translation metrics remain skeptical when text snippets are annotated with a score of 5 for being semantically analogous but syntactically the texts are expressed in a different form. For example, given the text snippets, *'There's not a lot you can do about that'* and *'I'm afraid there's not really a lot you can do'*, most machine translation metrics will not allocate full similarity scores due to the difference in lexical and stylistic ways in which the sentences are expressed.

Machine translation metrics' failure to capture similarity score extremes is evident in Figure 1 where there are no 0 and 5.0 predictions.

6 Conclusion

In this paper, we have described our submission to the English STS task for SemEval-2016. We have annotated the STS2012-2016 datasets with machine translation (MT) evaluation metric scores and trained a baseline linear regression and two tree ensemble models with the annotated data and achieved competitive results compared to the median pearson correlation scores from all participating systems.

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Main Conference and the Shared Task*, pages 385–393, Montréal, Canada.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 32–43, Atlanta, Georgia.

Eneko Agirre, Carmen Banea, Claire Cardic, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe.

2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Alberto Barrón-Cedeño, Lluís Màrquez, Maria Fuentes, Horacio Rodríguez, and Jordi Turmo. 2013. UPC-CORE: What Can Machine Translation Evaluation Metrics and Wikipedia Do for Estimating Semantic Textual Similarity? In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 143–147, Atlanta, Georgia.

Dario Bertero and Pascale Fung. 2015. Hltc-hkust: A neural network paraphrase classifier using translation metrics, semantic roles and lexical similarity features. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 23–28, Denver, Colorado, June.

Ergun Biçici and Josef van Genabith. 2013. CNGL-CORE: Referential Translation Machines for Measuring Semantic Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 234–240, Atlanta, Georgia.

Ergun Biçici and Andy Way. 2014. RTM-DCU: Referential Translation Machines for Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 487–496, Dublin, Ireland.

Ergun Bicici. 2015. Rtm-dcu: Predicting semantic similarity with referential translation machines. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 56–63, Denver, Colorado, June.

Alexandra Birch and Miles Osborne. 2010. Lrscore for evaluating lexical and reordering quality in mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 327–332.

Tianqi Chen and Carlos Guestrin. 2015. Xgboost: Reliable large-scale tree boosting system.

Tianqi Chen and Tong He. 2015. xgboost: extreme gradient boosting. *R package version 0.4-2*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Proceedings of the HLT: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253, Los Angeles, California.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of Twentieth International Joint Conference on Artificial Intelligence*.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Meritxell Gonzàlez, Alberto Barrón-Cedeo, and Llus Màrquez. 2014. Ipa and stout: Leveraging linguistic and source-based features for machine translation evaluation. In *Ninth Workshop on Statistical Machine Translation*, page 8.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal, September.

Nizar Habash and Ahmed Elkholy. 2008. Sepia: surface span extension to syntactic dependency precision-based mt evaluation. In *Proceedings of the NIST metrics for machine translation workshop at the association for machine translation in the Americas conference, AMTA-2008. Waikiki, HI*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Pingping Huang and Baobao Chang. 2014. SSMT:A Machine Translation Evaluation View To Paragraph-to-Sentence Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 585–589, Dublin, Ireland.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Cross level semantic similarity: an evaluation framework for universal measures of similarity. *Language Resources and Evaluation*, 50(1):5–33.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 605–612, Barcelona, Spain, July.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Steven Parker. 2008. Badger: A new machine translation metric. pages 21–25.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2012. UOW: Semantically Informed Text Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Main Conference and the Shared Task*, pages 673–678, Montréal, Canada.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 441–448. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809.

Miloš Stanojevic and Khalil Simaan. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. Dls@cu: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 148–153, Denver, Colorado, June.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China, July.

Liling Tan, Carolina Scarton, Lucia Specia, and Josef van Genabith. 2015. Usaar-sheffield: Semantic textual similarity with deep regression and machine translation evaluation metrics. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 85–89, Denver, Colorado, June.

Stephen Wan, Mark Dras, Robert Dale, and Cécile Paris. 2006. Using dependency-based features to take the para-farce out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop*, volume 2006.

Hao Zhang and Daniel Gildea. 2007. Factorization of synchronous context-free grammars in linear time. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 25–32.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.