The Impact of Multiword Expression Compositionality on Machine Translation Evaluation

Bahar Salehi,♠♣ Nitika Mathur,♠ Paul Cook♡ and Timothy Baldwin♠♣

♣ NICTA Victoria Research Laboratory
♠ Department of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia

♡ Faculty of Computer Science
University of New Brunswick
Fredericton, NB E3B 5A3, Canada

{bsalehi,nmathur}@student.unimelb.edu.au, paul.cook@unb.ca, tb@ldwin.net

Abstract

In this paper, we present the first attempt to integrate predicted compositionality scores of multiword expressions into automatic machine translation evaluation, in integrating compositionality scores for English noun compounds into the TESLA machine translation evaluation metric. The attempt is marginally successful, and we speculate on whether a larger-scale attempt is likely to have greater impact.

1 Introduction

While the explicit identification of multiword expressions ("MWEs": Sag et al. (2002), Baldwin and Kim (2009)) has been shown to be useful in various NLP applications (Ramisch, 2012), recent work has shown that automatic prediction of the degree of compositionality of MWEs also has utility, in applications including information retrieval ("IR": Acosta et al. (2011)) and machine translation ("MT": Weller et al. (2014), Carpuat and Diab (2010) and Venkatapathy and Joshi (2006)). For instance, Acosta et al. (2011) showed that by considering non-compositional MWEs as a single unit, the effectiveness of document ranking in an IR system improves, and Carpuat and Diab (2010) showed that by adding compositionality scores to the Moses SMT system (Koehn et al., 2007), they could improve translation quality.

This paper presents the first attempt to use MWE compositionality scores for the *evaluation* of MT system outputs. The basic intuition underlying this work is that we should sensitise the relative reward associated with partial mismatches between MT outputs and the reference translations, based on compositionality. For example, an MT output of *white tower* should not be rewarded for partial overlap with *ivory tower* in the reference translation, as *tower* here is most naturally interpreted compositionally in the MT output, but non-compositionally in the reference translation. On the other hand, a partial mismatch between *traffic signal* and *traffic light* should be rewarded, as the usage of *traffic* is highly compositional in both cases. That is, we ask the question: can we better judge the quality of translations if we have some means of automatically estimating the relative compositionality of MWEs, focusing on compound nouns, and the TESLA machine translation metric (Liu et al., 2010).

2 Related Work

In this section, we overview previous work on MT evaluation and measuring the compositionality of MWEs.

2.1 Machine Translation Evaluation

Automatic MT evaluation methods score MT system outputs based on similarity with reference translations provided by human translators. This scoring can be based on: (1) simple string similarity (Papineni et al., 2002; Snover et al., 2006); (2) shallow linguistic information such as lemmatisation, POS tagging and synonyms (Banerjee and Lavie, 2005; Liu et al., 2010); or (3) deeper linguistic information such as semantic roles (Giménez and Màrquez, 2008; Padó et al., 2009).

In this research, we focus on the TESLA MT eval-

Proceedings of NAACL-HLT 2015, pages 54–59,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

uation metric (Liu et al., 2010), which falls into the second group and uses a linear programming framework to automatically learn weights for matching n-grams of different types, making it easy to incorporate continuous-valued compositionality scores of MWEs.

2.2 Compositionality of MWEs

Earlier work on MWE compositionality (Bannard, 2006) approached the task via binary classification (compositional or non-compositional). However, there has recently been a shift towards regression analysis of the task, and prediction of a continuous-valued compositionality score (Reddy et al., 2011; Salehi and Cook, 2013; Salehi et al., 2014). This is the (primary) approach we take in this paper, as outlined in Section 3.2.

3 Methodology

3.1 Using compositionality scores in TESLA

In this section, we introduce TESLA and our method for integrating compositionality scores into the method.

Firstly, TESLA measures the similarity between the unigrams of the two given sentences (MT output and reference translation) based on the following three terms for each pairing of unigrams x and y:

$$\begin{aligned} S_{ms} &= \begin{cases} 1 & \text{if } lemma(x) = lemma(y) \\ \frac{a+b}{2} & \text{otherwise} \end{cases} \\ S_{lem}(x,y) &= I(lemma(x) = lemma(y)) \\ S_{pos}(x,y) &= I(POS(x) = POS(y)) \end{aligned}$$

where:

$$\begin{aligned} a &= I(synset(x) \cap synset(y)) \\ b &= I(POS(x) = POS(y)) \end{aligned}$$

$lemma$ returns the lemmatised unigram, POS returns the POS tag of the unigram, $synset$ returns the WordNet synsets associated with the unigram, and $I(.)$ is the indicator function.

The similarity between two n-grams $x = x^{1,2,\ldots,n}$ and $y = y^{1,2,\ldots,n}$ is measured as follows:

$$s(x,y) = \begin{cases} 0 & \text{if } \exists i, s(x^i, y^i) = 0 \\ \frac{1}{n}\sum_{i=1}^{n} s(x^i, y^i)) & \text{otherwise} \end{cases}$$

TESLA uses an integer linear program to find the phrase alignment that maximizes the similarity scores over the three terms (S_{ms}, S_{lem} and S_{pos}) for all n-grams.

In order to add the compositionality score to TESLA, we first identify MWEs in the MT output and reference translation. If an MWE in the reference translation aligns exactly with an MWE in the MT output, the weight remains as 1. Otherwise, we replace the computed weight computed for the noun compound with the product of computed weight and the compositionality degree of the MWE. This forces the system to be less flexible when encountering less compositional noun compounds. For instance, in TESLA, if the reference sentence contains *ivory tower* and the MT output contains *white building*, TESLA will align them with a score of 1. However, by multiplying this weight with the compositionality score (which should be very low for *ivory tower*), the alignment will have a much lower weight.

3.2 Predicting the compositionality of MWEs

In order to predict the compositionality of MWEs, we calculate the similarity between the MWE and each of its component words, using the three approaches detailed below. We calculate the overall compositionality of the MWE via linear interpolation over the component word scores, as:

$$\begin{aligned} comp(mwe) &= \alpha comp_c(mwe, w_1) + \\ &\quad (1-\alpha) comp_c(mwe, w_2) \end{aligned}$$

where mwe is, without loss of generality, made up of component words w_1 and w_2, and $comp_c$ is the compositionality score between mwe and the indicated component word. Based on the findings of Reddy et al. (2011), we set $\alpha = 0.7$.

Distributional Similarity (DS): the distributional similarity between the MWE and each of its components (Salehi et al., 2014), calculated based on cosine similarity over co-occurrence vectors, in the manner of Schütze (1997), using the 51st–1050th most frequent words in the corpus as dimensions. Context vectors were constructed from English Wikipedia.

	All sentences	Contains NC
METEOR	0.277	0.273
BLEU	0.216	0.206
TESLA	0.238	0.224
TESLA-DS	0.238	0.225
TESLA-SS+DS	0.238	0.225
TESLA-0/1	0.238	0.225

Table 1: Kendall's (τ) correlation over WMT 2013 (all-en), for the full dataset and also the subset of the data containing a noun compound in both the reference and the MT output

	All sentences	Contains NC
METEOR	0.436	0.500
BLEU	0.272	0.494
TESLA	0.303	0.467
TESLA-DS	0.305	0.464
TESLA-SS+DS	0.305	0.464
TESLA-0/1	0.308	0.464

Table 2: Pearson's (r) correlation results over the WMT all-en dataset, and the subset of the dataset that contains noun compounds

SS+DS: the arithmetic mean of DS and string similarity ("SS"), based on the findings of Salehi et al. (2014). SS is calculated for each component using the LCS-based string similarity between the MWE and each of its components in the original language as well as a number of translations (Salehi and Cook, 2013), under the hypothesis that compositional MWEs are more likely to be word-for-word translations in a given language than non-compositional MWEs. Following Salehi and Cook (2013), the translations were sourced from PanLex (Baldwin et al., 2010; Kamholz et al., 2014).

In Salehi and Cook (2013), the best translation languages are selected based on the training data. Since, we focus on NCs in this paper, we use the translation languages reported in that paper to work best for English noun compounds, namely: Czech, Norwegian, Portuguese, Thai, French, Chinese, Dutch, Romanian, Hindi and Russian.

4 Dataset

We evaluate our method over the data from WMT 2013, which is made up of a total of 3000 translations for five to-English language pairs (Bojar et al., 2013). As our judgements, we used: (1) the original pairwise preference judgements from WMT 2013 (i.e. which of translation A and B is better?); and (2) continuous-valued adequacy judgements for each MT output, as collected by Graham et al. (2014).

We used the Stanford CoreNLP parser (Klein and Manning, 2003) to identify English noun compounds in the translations. Among the 3000 sentences, 579 sentences contain at least one noun compound.

5 Results

We performed two evaluations, based on the two sets of judgements (pairwise preference or continuous-valued judgement for each MT output). In each case, we use three baselines (each applied at the segment level, meaning that individual sentences get a score): (1) METEOR (Banerjee and Lavie, 2005), (2) BLEU (Papineni et al., 2002), and (3) TESLA (without compositionality scores). We compare these with TESLA incorporating compositionality scores, based on DS ("TESLA-DS") and SS+DS ("TESLA-SS+DS"). We also include results for an exact match method which treats the MWEs as a single token, such that unless the MWE is translated exactly the same as in the reference translation, a score of zero results ("TESLA-0/1"). We did not experiment with the string similarity approach alone, because of the high number of missing translations in PanLex.

In the first experiment, we calculate the segment level Kendall's τ following the method used in the WMT 2013 shared task, as shown in Table 1, including the results over the subset of the data which contains a compound noun in both the reference and the MT output ("contains NC"). When comparing TESLA with and without MWE compositionality, we observe a tiny improvement with the inclusion of the compositionality scores (magnified slightly over the NC subset of the data), but not great enough to boost the score to that of METEOR. We also observe slightly lower correlations for TESLA-0/1 than TESLA-DS and TESLA-SS+DS, which consider degrees of compositionality, for fr-en, de-en and es-en (results not shown).

In the second experiment, we calculate Pearson's r correlation over the continuous-valued adequacy

Language Pair	*comp*	P→N	N→P	Δ
fr-en	DS	17	18	1
	SS+DS	14	16	2
	0/1	30	29	−1
de-en	DS	21	24	3
	SS+DS	14	18	4
	0/1	48	40	−8
es-en	DS	12	18	6
	SS+DS	11	17	6
	0/1	20	25	5
cs-en	DS	21	23	2
	SS+DS	14	16	2
	0/1	46	49	3
ru-en	DS	38	51	13
	SS+DS	29	39	10
	0/1	65	80	15

Table 3: The number of judgements that were ranked correctly by TESLA originally, but incorrectly with the incorporation of compositionality scores ("P→N") and vice versa ("N→P"), and the absolute improvement with compositionality scores ("Δ")

judgements, as shown in Table 2, again over the full dataset and also the subset of data containing compound nouns. The improvement here is slightly greater than for our first experiment, but not at a level of statistical significance (Graham and Baldwin, 2014). Perhaps surprisingly, the exact compositionality predictions produce a higher correlation than the continuous-valued compositionality predictions, but again, even with the inclusion of the compositionality features, TESLA is outperformed by METEOR. The correlation over the subset of the data containing compound nouns is markedly higher than that over the full dataset, but the r values with the inclusion of compositionality values are actually all slightly below those for the basic TESLA.

As a final analysis, we examine the relative impact on TESLA of the three compositionality methods, in terms of pairings of MT outputs where the ordering is reversed based on the revised TESLA scores. Table 3 details, for each language pairing, the number of pairwise judgements that were ranked correctly originally, but incorrectly when the compositionality score was incorporated ("P→N"); and also the number of pairwise judgements that were ranked incorrectly originally, and corrected with the incorporation of the compositionality judgements ("N→P").

Overall, the two compositionality methods perform better than the exact match method, and utilising compositionality has a more positive effect than negative. However, the difference between the numbers is, once again, very small, except for the ru-en language pair. The exact match method ("0/1") has a bigger impact, both positively and negatively, as a result of the polarisation of n-gram overlap scores for MWEs. We also noticed that the N→P sentences for SS+DS are a subset of the N→P sentences for DS. Moerover, the N→P sentences for DS are a subset of the N→P sentences for $0/1$; the same is true for the P→N sentences.

6 Discussion

As shown in the previous section, the incorporation of compositionality scores can improve the quality of MT evaluation based on TESLA. However, the improvements are very small and not statistically significant. Part of the reason is that we focus exclusively on noun compounds, which are contiguous and relatively easy to translate for MT systems (Koehn and Knight, 2003). Having said that, preliminary error analysis would suggest that most MT systems have difficulty translating non-compositional noun compounds, although then again, most noun compounds in the WMT 2013 shared task are highly compositional, limiting the impact of compositionality scores. We speculate that, for the method to have greater impact, we would need to target a larger set of MWEs, including non-contiguous MWEs such as split verb particle constructions (Kim and Baldwin, 2010).

Further error analysis suggests that incorrect identification of noun compounds in a reference sentence can have a negative impact on MT evaluation. For example, *year student* is mistakenly identified as an MWE in ... *a 21-year-old final year student at Temple* ....

Furthermore, when an MWE occurs in a reference translation, but not an MT system's output, incorporating the compositionality score can sometimes result in an error. For instance, in the first example in Table 4, the reference translation contains the compound noun *cash flow*. According to the dataset, the output of MT system 1 is better than that of MT sys-

Reference	This means they are much better for our cash flow.
MT system 1	That is why they are for our money flow of a much better.
MT system 2	Therefore, for our cash flow much better.
Reference	‘I felt like I was in a luxury store,’ he recalls.
MT system 1	‘I feel as though I am in a luxury trade,’ recalls soldier.
MT system 2	‘I felt like a luxury in the store,’ he recalled the soldier.

Table 4: Two examples from the all-en dataset. Each example shows a reference translation, and the outputs of two machine translation systems. In each case, the output of MT system 1 is annotated as the better translation.

tem 2. However, since the former translation does not contain an exact match for *cash flow*, our method decreases the alignment score by multiplying it by the compositionality score for *cash flow*. As a result, the overall score for the first translation becomes less than that of the second, and our method incorrectly chooses the latter as a better translation.

Incorrect estimation of compositionality scores can also have a negative effect on MT evaluation. In the second example in Table 4, the similarity score between *luxury store* and *luxury trade* given by TESLA is 0.75. The compositionality score, however, is estimated as 0.22. The updated similarity between *luxury trade* and *luxury store* is therefore 0.16, which in this case results in our method incorrectly selecting the second sentence as the better translation.

7 Conclusion

This paper described the first attempt at integrating MWE compositionality scores into an automatic MT evaluation metric. Our results show a marginal improvement with the incorporation of compositionality scores of noun compounds.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions. NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme.

References

Otavio Acosta, Aline Villavicencio, and Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 101–109, Portland, USA.

Timothy Baldwin and Su Nam Kim. 2009. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, Boca Raton, USA, 2nd edition.

Timothy Baldwin, Jonathan Pool, and Susan M. Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 37–40, Beijing, China.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Colin James Bannard. 2006. *Acquiring Phrasal Lexicons from Corpora*. Ph.D. thesis, University of Edinburgh.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, USA.

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on*

Empirical Methods in Natural Language Processing (EMNLP 2014), pages 172–176, Doha, Qatar.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the EACL (EACL 2014)*, pages 443–451, Gothenburg, Sweden.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland.

Su Nam Kim and Timothy Baldwin. 2010. How to pick out token instances of English verb-particle constructions. *Language Resources and Evaluation*, 44(1-2):97–113.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, pages 3–10, Whistler, Canada.

Philipp Koehn and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. Tesla: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 354–359.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Chris Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 297–305.

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 (ACL 2002)*, pages 311–318.

Carlos Ramisch. 2012. A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, pages 61–66, Jeju Island, Korea.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP*, pages 210–218, Chiang Mai, Thailand.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing Computational Linguistics (CICLing-2002)*, pages 189–206, Mexico City, Mexico.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 266–275, Atlanta, Georgia, USA, June.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden, April.

Hinrich Schütze. 1997. *Ambiguity Resolution in Language Learning*. CSLI Publications, Stanford, USA.

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*, pages 223–231.

Sriram Venkatapathy and Aravind K Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20–27.

Marion Weller, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Distinguishing degrees of compositionality in compound splitting for statistical machine translation. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*, pages 81–90, Dublin, Ireland.