Lexical Access Preference and Constraint Strategies for Improving Multiword Expression Association within Semantic MT Evaluation

Dekai Wu   **Lo Chi-kiu**   **Markus Saers**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{dekai|jackielo|masaers|dekai}@cs.ust.hk

Abstract

We examine lexical access preferences and constraints in computing multiword expression associations from the standpoint of a high-impact extrinsic task-based performance measure, namely semantic machine translation evaluation. In automated MT evaluation metrics, machine translations are compared against human reference translations, which are almost never worded exactly the same way except in the most trivial of cases. Because of this, one of the most important factors in correctly predicting semantic translation adequacy is the accuracy of recognizing alternative lexical realizations of the same multiword expressions in semantic role fillers. Our results comparing bag-of-words, maximum alignment, and inversion transduction grammars indicate that cognitively motivated ITGs provide superior lexical access characteristics for multiword expression associations, leading to state-of-the-art improvements in correlation with human adequacy judgments.

1 Introduction

We investigate lexical access strategies in the context of computing multiword expression associations within automatic semantic MT evaluation metrics—a high-impact real-world extrinsic task-based performance measure. The inadequacy of lexical coverage of multiword expressions is one of the serious issues in machine translation and automatic MT evaluation; there are simply too many forms to enumerate explicitly within the lexicon. Automatic MT evaluation has driven machine translation research for a decade and a half, but until recently little has been done to use lexical semantics as the main foundation for MT metrics. Common surface-form oriented metrics like BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006) do not explicitly reflect semantic similarity between the reference and machine translations. Several large scale meta-evaluations (Callison-Burch *et al.*, 2006; Koehn and Monz, 2006) have in fact reported that BLEU significantly disagrees with human judgments of translation adequacy.

Recently, the MEANT semantic frame based MT evaluation metrics (Lo and Wu, 2011a, 2012; Lo *et al.*, 2012; Lo and Wu, 2013b), have instead directly couched MT evaluation in the more cognitive terms of semantic frames, by measuring the degree to which the basic event structure is preserved by translation—the "who did what to whom, for whom, when, where, how and why" (Pradhan *et al.*, 2004)—emphasizing that a good translation is one that can successfully be understood by a human. Across a variety of language pairs and genres, MEANT was shown to correlate better with human adequacy judgment than both n-gram based MT evaluation metrics such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), and METEOR (Banerjee and Lavie, 2005), as well as edit-distance based metrics such as CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006) when evaluating MT output (Lo and Wu, 2011a, 2012; Lo *et al.*, 2012; Lo and Wu, 2013b; Macháček and Bojar, 2013). Furthermore, tuning the parameters of MT systems with MEANT instead of BLEU or TER robustly improves translation

Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon, pages 144–153,
Dublin, Ireland, August 23, 2014.

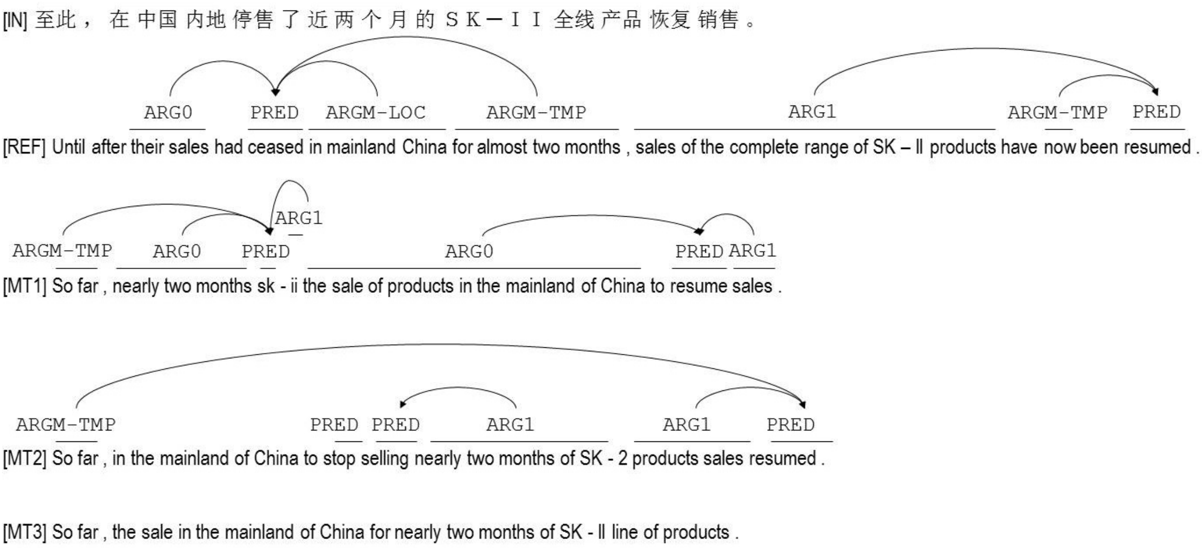

Figure 1: Examples of automatic shallow semantic parses. Both the reference and machine translations are parsed using automatic English SRL. There are no semantic frames for MT3 since automatic SRL decided to drop the predicate.

adequacy (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b) across different languages (English and Chinese) and different genres (formal newswire text, informal web forum text and informal public speech).

Because of this, we have chosen to run our lexical association experiments in the context of the necessity of recognizing matching semantic role fillers, approximately 85% of which are multiword expressions in our data, the overwhelming majority of which would not be enumerated within conventional lexicons. We compare four common lexical access approaches to aggregation, preferences, and constraints: bag-of-words, two different types of maximal alignment, and inversion transduction grammar based methods.

2 Background

The MEANT metric measures weighted f-scores over corresponding semantic frames and role fillers in the reference and machine translations. Whereas HMEANT uses human annotation, the automatic versions of MEANT instead replace humans with automatic SRL and alignment algorithms. MEANT typically outperforms BLEU, NIST, METEOR, WER, CDER and TER in correlation with human adequacy judgment, and is relatively easy to port to other languages, requiring only an automatic semantic parser and a monolingual corpus of the output language, which is used to gauge lexical similarity between the semantic role fillers of the reference and translation. More precisely, MEANT computes scores as follows:

1. Apply an automatic shallow semantic parser to both the references and MT output. (Figure 1 shows examples of automatic shallow semantic parses on both reference and MT.)

2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between the references and MT output according to the lexical similarities of the predicates.

3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between the reference and MT output according to the lexical similarity of role fillers.

4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers according to the following definitions:

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG j of aligned frame i in MT} \\
q^1_{i,j} &\equiv \text{ARG j of aligned frame i in REF} \\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}} \\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG j} \\
\mathbf{e}_{i,\text{pred}} &\equiv \text{the pred of the aligned frame } i \textit{ of the machine translation} \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the pred of the aligned frame } i \textit{ of the reference translation}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{e}_{i,j} &\equiv \text{the ARG } j \text{ of the aligned frame } i \textit{ of the machine translation} \\
\mathbf{f}_{i,j} &\equiv \text{the ARG } j \text{ of the aligned frame } i \textit{ of the reference translation} \\
s(e,f) &= \text{lexical similarity of token } e \text{ and } f \\
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e\in\mathbf{e}} \max\limits_{f\in\mathbf{f}} s(e,f)}{|\,\mathbf{e}\,|} \\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f\in\mathbf{f}} \max\limits_{e\in\mathbf{e}} s(e,f)}{|\,\mathbf{f}\,|} \\
\text{precision} &= \frac{\sum_i w^0_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^0_{i,j}|}}{\sum_i w^0_i} \\
\text{recall} &= \frac{\sum_i w^1_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^1_{i,j}|}}{\sum_i w^1_i} \\
\text{MEANT} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\end{aligned}
$$

where the possible approaches to defining the lexical associations $s_{i,\text{pred}}$ and $s_{i,j}$ are discussed in the following section. $q^0_{i,j}$ and $q^1_{i,j}$ are the argument of type j in frame i in MT and REF, respectively. w^0_i and w^1_i are the weights for frame i in MT and REF, respectively. These weights estimate the degree of contribution of each frame to the overall meaning of the sentence. w_{pred} and w_j are the weights of the lexical similarities of the predicates and role fillers of the arguments of type j of all frame between the reference translations and the MT output. There is a total of 12 weights for the set of semantic role labels in MEANT as defined in Lo and Wu (2011b). For MEANT, they are determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments (Lo and Wu, 2011a). For UMEANT (Lo and Wu, 2012), they are estimated in an unsupervised manner using relative frequency of each semantic role label in the references and thus UMEANT is useful when human judgments on adequacy of the development set are unavailable.

3 Comparison of multiword expression association approaches

To assess alternative lexical access preferences and constraints for computing multiword expression associations, we now consider four alternative approaches to defining the lexical similarities $s_{i,\text{pred}}$ and $s_{i,j}$, all of which employ a standard context vector model of the individual words/tokens in the multiword expression arguments between the reference and machine translations, as descibed by Lo *et al.* (2012) and Tumuluru *et al.* (2012).

3.1 Bag of words (geometric mean)

The original MEANT approaches employed standard a bag-of-words strategy for lexical association. This baseline approach applies no alignment constraints on multiword expressions:

$$
\begin{aligned}
s_{i,\text{pred}} &= e^{\frac{\sum_{e\in\mathbf{e}_{i,\text{pred}}} \sum_{f\in\mathbf{f}_{i,\text{pred}}} \lg(s(e,f))}{|\mathbf{e}_{i,\text{pred}}|\cdot|\mathbf{f}_{i,\text{pred}}|}} \\
s_{i,j} &= e^{\frac{\sum_{e\in\mathbf{e}_{i,j}} \sum_{f\in\mathbf{f}_{i,j}} \lg(s(e,f))}{|\mathbf{e}_{i,j}|\cdot|\mathbf{f}_{i,j}|}}
\end{aligned}
$$

3.2 Maximum alignment (precision-recall average)

In the first maximum alignment based approach we will consider, the definitions of $s_{i,\text{pred}}$ and $s_{i,j}$ are inspired by Mihalcea *et al.* (2006) who normalize phrasal similarities according to the phrase length.

$$\begin{aligned} s_{i,\text{pred}} &= \frac{1}{2}(\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}) \\ s_{i,j} &= \frac{1}{2}(\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}) \end{aligned}$$

3.3 Maximum alignment (f-score)

The second of the maximum alignment based approaches replaces the above linear averaging of precision and recall with a proper f-score. Although this is less consistent with the previous literature, such as Mihalcea *et al.* (2006), it seems more consistent with the overall f-score based approach of MEANT, and thus we include it in our comparison as a variant of the maximum alignment strategy.

$$\begin{aligned} s_{i,\text{pred}} &= \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}} \\ s_{i,j} &= \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}} \end{aligned}$$

3.4 Inversion transduction grammar based

There has been to date relatively little use of inversion transduction grammars (Wu, 1997) to improve the accuracy of MT evaluation metrics—despite (1) long empirical evidence the vast majority of translation patterns between human languages can be accommodated within ITG constraints, and (2) the observation that most current state-of-the-art SMT systems employ ITG decoders. Especially when considering *semantic* MT metrics, ITGs would seem to be a natural strategy for multiword expression association for several cognitively motivated reasons, having to do with language universal properties of cross-linguistic semantic frame structure.

To begin with, it is quite natural to think of sentences as having been generated from an abstract concept using a rewriting system: a stochastic grammar predicts how frequently any particular realization of the abstract concept will be generated. The bilingual analogy is a *transduction grammar* generating *a pair* of possible realizations of *the same* underlying concept. Stochastic transduction grammars predict how frequently a particular pair of realizations will be generated, and thus represent a good way to evaluate how well a pair of sentences correspond to each other.

The particular class of transduction grammars known as ITGs tackle the problem that the (bi)parsing complexity for general **syntax-directed transductions** (Aho and Ullman, 1972) is exponential. By constraining a syntax-directed transduction grammar to allow only monotonic **straight** and **inverted** reorderings, or equivalently permitting only binary or ternary rank rules, it is possible to isolate the low end of that hierarchy into a single equivalence class of **inversion transductions**. ITGs are guaranteed to have a two-normal form similar to context-free grammars, and can be biparsed in polynomial time and space ($O\left(n^6\right)$ time and $O\left(n^4\right)$ space). It is also possible to do approximate biparsing in $O\left(n^3\right)$ time (Saers *et al.*, 2009). These polynomial complexities makes it feasible to estimate the parameters of an ITG using standard machine learning techniques such as expectation maximization (Wu, 1995b) .

At the same time, inversion transductions have also been directly shown to be more than sufficient to account for the reordering that occur within semantic frame alternations (Addanki *et al.*, 2012). This language universal property has an evolutionary explanation in terms of computational efficiency and cognitive load for language learnability and interpretability (Wu, 2014).

ITGs are thus an appealing alternative for evaluating the possible links between both semantic role fillers in different languages as well as the predicates, and how these parts fit together to form entire semantic frames. We believe that ITGs are not only capable of generating the desired structural correspondences between the semantic structures of two languages, but also provide meaningful constraints to prevent alignments from wandering off in the wrong direction.

Following this reasoning, alternate definitions of $s_{i,\text{pred}}$ and $s_{i,j}$ can be constructed in terms of bracketing ITGs (also known as BITGs or BTGs) which are ITGs containing only a single non-differentiated nonterminal category (Wu, 1995a). The idea is to attack a potential weakness of the foregoing three lexical association strategies, namely that word/token alignments between the reference and machine translations are severely underconstrained. No bijectivity or permutation restrictions are applied, even between compositional segments where this should be natural. This can cause multiword expressions of semantic role fillers to be matched even when they should not be. In contrast, using a bracketing inversion transduction grammar can potentially better constrain permissible token alignment patterns between aligned role filler phrases. Figure 2 illustrates how the ITG constraints are consistent with the needed permutations between semantic role fillers across the reference and machine translations for a sample sentence from the evaluation data.

In this approach, both alignment and scoring are performed utilizing a length-normalized weighted BITG (Wu, 1997; Zens and Ney, 2003; Saers and Wu, 2009; Addanki *et al.*, 2012). We define $s_{i,\text{pred}}$ and $s_{i,j}$ as follows.

$$
\begin{aligned}
s_{i,\text{pred}} &= \lg^{-1}\left(\frac{\lg\left(P\left(\mathrm{A} \stackrel{*}{\Rightarrow} \mathbf{e}_{i,\text{pred}}/\mathbf{f}_{i,\text{pred}}|G\right)\right)}{\max(|\ \mathbf{e}_{i,\text{pred}}\ |, |\ \mathbf{f}_{i,\text{pred}}\ |)}\right) \\
s_{i,j} &= \lg^{-1}\left(\frac{\lg\left(P\left(\mathrm{A} \stackrel{*}{\Rightarrow} \mathbf{e}_{i,j}/\mathbf{f}_{i,j}|G\right)\right)}{\max(|\ \mathbf{e}_{i,j}\ |, |\ \mathbf{f}_{i,j}\ |)}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
G &\equiv \langle\{\mathrm{A}\}, \mathcal{W}^0, \mathcal{W}^1, \mathcal{R}, \mathrm{A}\rangle \\
\mathcal{R} &\equiv \{\mathrm{A} \rightarrow [\mathrm{AA}], \mathrm{A} \rightarrow \langle\mathrm{AA}\rangle, \mathrm{A} \rightarrow e/f\}
\end{aligned}
$$

$$
\begin{aligned}
p([\mathrm{AA}]|\mathrm{A}) &= p(\langle\mathrm{AA}\rangle|\mathrm{A}) = 1 \\
p(e/f|\mathrm{A}) &= s(e,f)
\end{aligned}
$$

Here G is a bracketing ITG whose only nonterminal is A, and $\mathcal{R}$ is a set of transduction rules with $e \in \mathcal{W}^0 \cup \{\epsilon\}$ denoting a token in the MT output (or the *null* token) and $f \in \mathcal{W}^1 \cup \{\epsilon\}$ denoting a token in the reference translation (or the *null* token). The rule probability (or more accurately, rule weight) function p is set to be 1 for structural transduction rules, and for lexical transduction rules it is defined by MEANT's lexical similarity measure on English Gigaword context vectors. To calculate the inside probability (or more accurately, inside score) of a pair of segments, $P\left(\mathrm{A} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}|G\right)$, we use the algorithm described in Saers *et al.* (2009). Given this, $s_{i,\text{pred}}$ and $s_{i,j}$ now represent the length normalized BITG parse scores of the predicates and role fillers of the arguments of type j between the reference and machine translations.

4 Experiments

In this section we discuss experiments comparing the four alternative lexical access preference and constraint strategies.

4.1 Experimental setup

We compared using the DARPA GALE P2.5 Chinese-English translation test set, as used in Lo and Wu (2011a). The corpus includes the Chinese input sentences, each accompanied by an English reference translation and three participating state-of-the-art MT systems' output.

We computed sentence-level correlations following the benchmark assessment procedure used by WMT and NIST MetricsMaTr (Callison-Burch *et al.*, 2008, 2010, 2011, 2012; Macháček and Bojar, 2013), which use Kendall's τ correlation coefficient, to evaluate the correlation of evaluation metrics against human judgment on ranking the translation adequacy of the three systems' output. A higher value for Kendall's τ indicates more similarity to the human adequacy rankings by the evaluation metrics. The range of possible values of Kendall's τ correlation coefficient is [-1, 1], where 1 means the

Table 1: Sentence-level correlation with human adequacy judgements on different partitions of GALE P2.5 data. For reference, the human HMEANT upper bound is 0.53—so the fully automatic ITG based MEANT approximation is not far from closing the gap.

	Kendall correlation
MEANT + ITG based	**0.51**
MEANT + maximum alignment (f-score)	0.48
MEANT + maximum alignment (average of precision & recall)	0.46
MEANT + bag of words (geometric mean)	0.38
NIST	0.29
METEOR	0.20
BLEU	0.20
TER	0.20
PER	0.20
CDER	0.12
WER	0.10

systems are ranked in the same order as the human judgment by the evaluation metric; and -1 means the systems are ranked in the reverse order as human judgment by the evaluation metric.

For both reference and machine translations, the ASSERT (Pradhan *et al.*, 2004) semantic role labeler was used to automatically predict semantic parses.

4.2 Results and discussion

The sentence-level correlations in Table 1 show that the ITG based strategy outperforms other automatic metrics in correlation with human adequacy judgment. Note that this was achieved with no tuning whatsoever of the rule weights (suggesting that the performance could be further improved in the future by slightly optimizing the ITG weights).

The ITG based strategy shows 3 points improvement over the next best strategy, which is maximal alignment under f-score aggregation. The ITG based approach produces much higher HAJ correlations than any of the other metrics.

In fact, the ITG based strategy even comes within a few points of the human upper bound benchmark HAJ correlations computed using the human labeled semantic frames and alignments used in the HMEANT.

Data analysis reveals two reasons that the ITG based strategy correlates with human adequacy judgement more closely than the other approaches. First, BITG constraints indeed provide more accurate phrasal similarity aggregation, compared to the naive bag-of-words based heuristics. Similar results have been observed while trying to estimate word alignment probabilities where BITG constraints outperformed alignments from GIZA++ (Saers and Wu, 2009). Secondly, the permutation and bijectivity constraints enforced by the ITG provide better leverage to reject token alignments when they are not appropriate, compared with the maximal alignment approach which tends to be rather promiscuous. The ITG tends whenever appropriate to accept clean, sparse alignments for role fillers, prefering to leave tokens unaligned instead of aligning them anyway as the other strategies tend to do. Note that it is not simply a matter of lowering thresholds for accepting token alignments: Tumuluru *et al.* (2012) showed that the competitive linking approach (Melamed, 1996) does not work as well as the strategies considered in this paper, whereas the ITG appears to be selective about the token alignments in a manner that better fits the semantic structure.

5 Conclusion

We have compared four alternative lexical access strategies for aggregation, preferences, and constraints in scoring multiword expression associations that are far too numerous to be explicitly enumerated in lexicons, within the context of semantic frame based machine translation evaluation: bag-of-words,

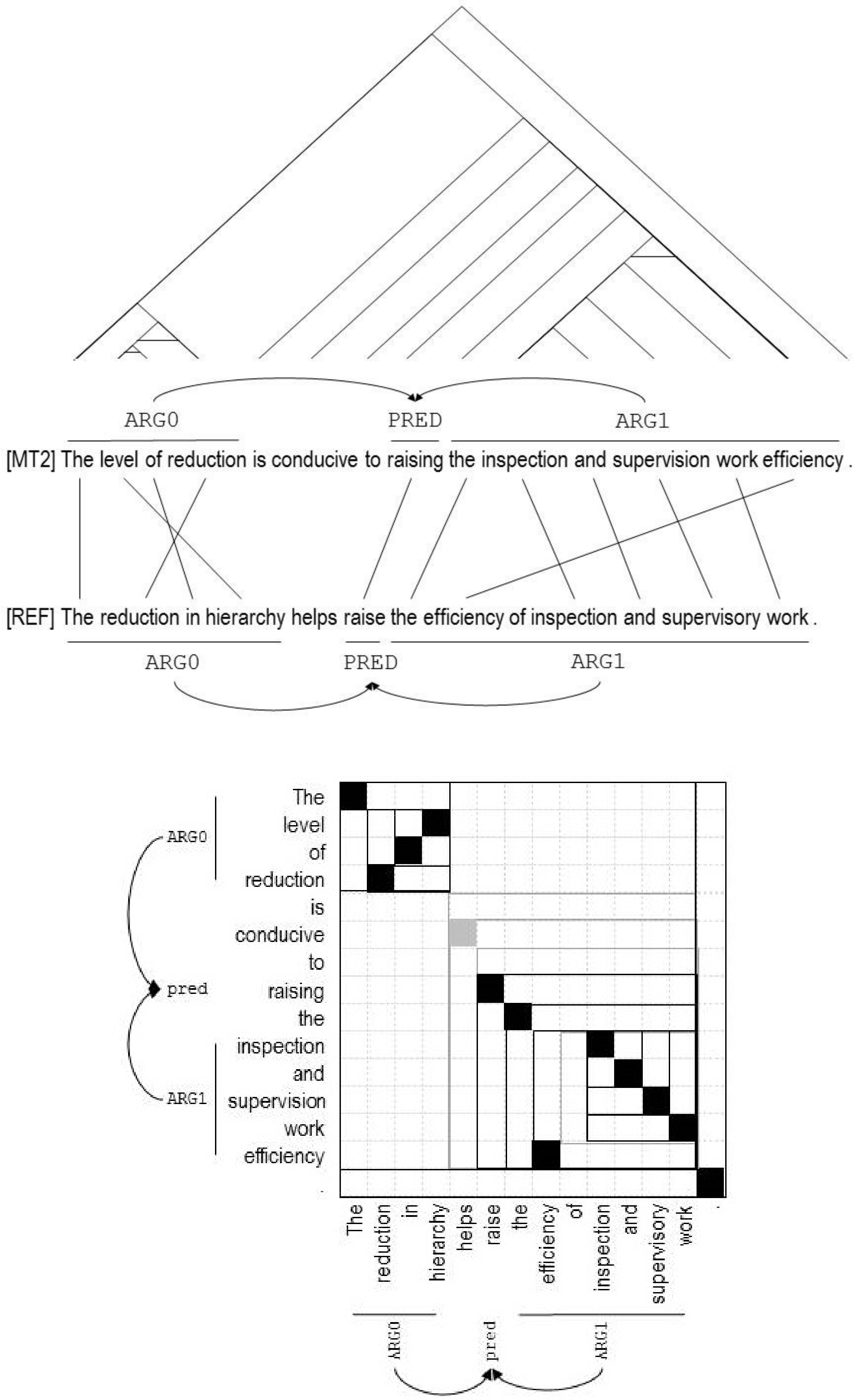

Figure 2: An example of aligning automatic shallow semantic parses under ITGs, visualized using both biparse tree and alignment matrix depictions, for the Chinese input sentence 层级的减少有利于提高检查监督工作的效率。 Both the reference and machine translations are parsed using automatic English SRL. Compositional alignments between the semantic frames and the tokens within role filler phrases obey inversion transduction grammars.

two maximum alignment based approaches, and an inversion transduction grammar based approach. Controlled experiments within the MEANT semantic MT evaluation framework shows that the cognitively motivated ITG based strategy achieves significantly higher correlation with human adequacy judgments of MT output quality than the more typically used lexical association approaches. The results show how to improve upon previous research showing that MEANT's explicit use of semantic frames leads to state-of-the-art automatic MT evaluation, by aligning and scoring semantic frames under a simple, consistent ITG that provides empirically informative permutation and bijectivity biases, instead of more naive maximal alignment or bag-of-words assumptions.

Cognitive studies of the lexicon are often described using intrinsic measures of quality. Our experiments complement this by situating the empirical comparisons within extrinsic real-world task-based performance measures. We believe that progress can be accelerated via a combination of intrinsic and extrinsic measures of lexicon acquisition and access models.

Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract nos. HR0011-12-C-0014 and HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008, and GRF612806. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC.

References

Karteek Addanki, Chi-kiu Lo, Markus Saers, and Dekai Wu. LTG vs. ITG coverage of cross-lingual verb frame alternations. In *16th Annual Conference of the European Association for Machine Translation (EAMT-2012)*, Trento, Italy, May 2012.

Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation, and Compiling*. Prentice-Halll, Englewood Cliffs, New Jersey, 1972.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, June 2005.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Third Workshop on Statistical Machine Translation (WMT-08)*, 2008.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Pryzbocki, and Omar Zaidan. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR (WMT10)*, pages 17–53, Uppsala, Sweden, 15-16 July 2010.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan. Findings of the 2011 Workshop on Statistical Machine Translation. In *6th Workshop on Statistical Machine Translation (WMT 2011)*, 2011.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, pages 10–51, 2012.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *The second international conference on Human Language Technology Research (HLT '02)*, San Diego, California, 2002.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Workshop on Statistical Machine Translation (WMT-06)*, 2006.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDer: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

Chi-kiu Lo and Dekai Wu. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, 2011.

Chi-kiu Lo and Dekai Wu. SMT vs. AI redux: How semantic frames evaluate MT more accurately. In *Twenty-second International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2011.

Chi-kiu Lo and Dekai Wu. Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In *Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, 2012.

Chi-kiu Lo and Dekai Wu. Can informal genres be better translated by tuning on automatic semantic metrics? In *14th Machine Translation Summit (MT Summit XIV)*, 2013.

Chi-kiu Lo and Dekai Wu. MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based mt evaluation metric. In *8th Workshop on Statistical Machine Translation (WMT 2013)*, 2013.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. Fully automatic semantic MT evaluation. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu. Improving machine translation by training against an automatic semantic frame based evaluation metric. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

Chi-kiu Lo, Meriem Beloucif, and Dekai Wu. Improving machine translation into Chinese by tuning against Chinese MEANT. In *International Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

Matouš Macháček and Ondřej Bojar. Results of the WMT13 metrics shared task. In *Eighth Workshop on Statistical Machine Translation (WMT 2013)*, Sofia, Bulgaria, August 2013.

I. Dan Melamed. Automatic construction of clean broad-coverage translation lexicons. In *2nd Conference of the Association for Machine Translation in the Americas (AMTA-1996)*, 1996.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *The Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, volume 21. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. A evaluation tool for machine translation: Fast evaluation for MT research. In *The Second International Conference on Language Resources and Evaluation (LREC 2000)*, 2000.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, Pennsylvania, July 2002.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow semantic parsing using support vector machines. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, 2004.

Markus Saers and Dekai Wu. Improving phrase-based translation via word alignments from stochastic inversion transduction grammars. In *Third Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, pages 28–36, Boulder, Colorado, June 2009.

Markus Saers, Joakim Nivre, and Dekai Wu. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *11th International Conference on Parsing Technologies (IWPT'09)*, pages 29–32, Paris, France, October 2009.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *7th Biennial Conference Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, August 2006.

Anand Karthik Tumuluru, Chi-kiu Lo, and Dekai Wu. Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic MT evaluation. In *26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 26)*, 2012.

Dekai Wu. An algorithm for simultaneously bracketing parallel texts by aligning words. In *33rd Annual Meeting of the Association for Computational Linguistics (ACL 95)*, pages 244–251, Cambridge, Massachusetts, June 1995.

Dekai Wu. Trainable coarse bilingual grammars for parallel text bracketing. In *Third Annual Workshop on Very Large Corpora (WVLC-3)*, pages 69–81, Cambridge, Massachusetts, June 1995.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, 1997.

Dekai Wu. The magic number 4: Evolutionary pressures on semantic frame structure. In *10th International Conference on the Evolution of Language (Evolang X)*, Vienna, Apr 2014.

Richard Zens and Hermann Ney. A comparative study on reordering constraints in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 144–151, Stroudsburg, Pennsylvania, 2003.