Source-Language Entailment Modeling for Translating Unknown Terms

Shachar Mirkin[§], **Lucia Specia**[†], **Nicola Cancedda**[†], **Ido Dagan**[§], **Marc Dymetman**[†], **Idan Szpektor**[§]
§ Computer Science Department, Bar-Ilan University
† Xerox Research Centre Europe
{mirkins,dagan,szpekti}@cs.biu.ac.il
{lucia.specia,nicola.cancedda,marc.dymetman}@xrce.xerox.com

Abstract

This paper addresses the task of handling unknown terms in SMT. We propose using source-language monolingual models and resources to paraphrase the source text prior to translation. We further present a conceptual extension to prior work by allowing translations of *entailed* texts rather than paraphrases only. A method for performing this process efficiently is presented and applied to some 2500 sentences with unknown terms. Our experiments show that the proposed approach substantially increases the number of properly translated texts.

1 Introduction

Machine Translation systems frequently encounter terms they are not able to translate due to some missing knowledge. For instance, a Statistical Machine Translation (SMT) system translating the sentence "*Cisco filed a lawsuit against Apple for patent violation*" may lack words like *filed* and *lawsuit* in its phrase table. The problem is especially severe for languages for which parallel corpora are scarce, or in the common scenario when the SMT system is used to translate texts of a domain different from the one it was trained on.

A previously suggested solution (Callison-Burch et al., 2006) is to learn paraphrases of source terms from multilingual (parallel) corpora, and expand the phrase table with these paraphrases[1]. Such solutions could potentially yield a paraphrased sentence like "*Cisco sued Apple for patent violation*", although their dependence on bilingual resources limits their utility.

In this paper we propose an approach that consists in directly replacing unknown source terms, using source-language resources and models in order to achieve two goals.

The first goal is coverage increase. The availability of bilingual corpora, from which paraphrases can be learnt, is in many cases limited. On the other hand, monolingual resources and methods for extracting paraphrases from monolingual corpora are more readily available. These include manually constructed resources, such as WordNet (Fellbaum, 1998), and automatic methods for paraphrases acquisition, such as DIRT (Lin and Pantel, 2001). However, such resources have not been applied yet to the problem of substituting unknown terms in SMT. We suggest that by using such monolingual resources we could provide paraphrases for a larger number of texts with unknown terms, thus increasing the overall coverage of the SMT system, i.e. the number of texts it properly translates.

Even with larger paraphrase resources, we may encounter texts in which not all unknown terms are successfully handled through paraphrasing, which often results in poor translations (see Section 2.1). To further increase coverage, we therefore propose to generate and translate texts that convey a somewhat more general meaning than the original source text. For example, using such approach, the following text could be generated: "*Cisco accused Apple of patent violation*". Although less informative than the original, a translation for such texts may be useful. Such non-symmetric relationships (as between *filed a lawsuit* and *accused*) are difficult to learn from parallel corpora and therefore monolingual resources are more appropriate for this purpose.

The second goal we wish to accomplish by employing source-language resources is to rank the alternative generated texts. This goal can be achieved by using context-models on the source language prior to translation. This has two advantages. First, the ranking allows us to prune some

[1] As common in the literature, we use the term *paraphrases* to refer to texts of equivalent meaning, of any length from single words (synonyms) up to complete sentences.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 791–799,
Suntec, Singapore, 2-7 August 2009.

candidates before supplying them to the translation engine, thus improving translation efficiency. Second, the ranking may be combined with target language information in order to choose the best translation, thus improving translation quality.

We position the problem of generating alternative texts for translation within the Textual Entailment (TE) framework (Giampiccolo et al., 2007). TE provides a generic way for handling language variability, identifying when the meaning of one text is *entailed* by the other (i.e. the meaning of the entailed text can be inferred from the meaning of the entailing one). When the meanings of two texts are equivalent (paraphrase), entailment is mutual. Typically, a more general version of a certain text is entailed by it. Hence, through TE we can formalize the generation of both equivalent and more general texts for the source text. When possible, a paraphrase is used. Otherwise, an alternative text whose meaning is entailed by the original source is generated and translated.

We assess our approach by applying an SMT system to a text domain that is different from the one used to train the system. We use WordNet as a source language resource for entailment relationships and several common statistical context-models for selecting the best generated texts to be sent to translation. We show that the use of source language resources, and in particular the extension to non-symmetric textual entailment relationships, is useful for substantially increasing the amount of texts that are properly translated. This increase is observed relative to both using paraphrases produced by the same resource (WordNet) and using paraphrases produced from multilingual parallel corpora. We demonstrate that by using simple context-models on the source, efficiency can be improved, while translation quality is maintained. We believe that with the use of more sophisticated context-models further quality improvement can be achieved.

2 Background

2.1 Unknown Terms

A very common problem faced by machine translation systems is the need to translate terms (words or multi-word expressions) that are not found in the system's lexicon or phrase table. The reasons for such unknown terms in SMT systems include scarcity of training material and the application of the system to text domains that differ from the ones used for training.

In SMT, when unknown terms are found in the source text, the systems usually omit or copy them literally into the target. Though copying the source words can be of some help to the reader if the unknown word has a cognate in the target language, this will not happen in the most general scenario where, for instance, languages use different scripts. In addition, the presence of a single unknown term often affects the translation of wider portions of text, inducing errors in both lexical selection and ordering. This phenomenon is demonstrated in the following sentences, where the translation of the English sentence (1) is acceptable only when the unknown word (in bold) is replaced with a translatable paraphrase (3):

1. "*..., despite bearing the heavy burden of the unemployed 10% or more of the* ***labor*** *force.*"
2. "*..., malgré la lourde charge de compte le 10% ou plus de chômeurs* ***labor*** *la force .*"
3. "*..., malgré la lourde charge des chômeurs de 10% ou plus de la force du* ***travail****.*"

Several approaches have been proposed to deal with unknown terms in SMT systems, rather than omitting or copying the terms. For example, (Eck et al., 2008) replace the unknown terms in the source text by their definition in a monolingual dictionary, which can be useful for gisting. To translate across languages with different alphabets approaches such as (Knight and Graehl, 1997; Habash, 2008) use transliteration techniques to tackle proper nouns and technical terms. For translation from highly inflected languages, certain approaches rely on some form of lexical approximation or morphological analysis (Koehn and Knight, 2003; Yang and Kirchhoff, 2006; Langlais and Patry, 2007; Arora et al., 2008). Although these strategies yield gain in coverage and translation quality, they only account for unknown terms that should be transliterated or are variations of known ones.

2.2 Paraphrasing in MT

A recent strategy to broadly deal with the problem of unknown terms is to paraphrase the source text with terms whose translation is known to the system, using paraphrases learnt from multilingual corpora, typically involving at least one "pivot" language different from the target language of immediate interest (Callison-Burch et

al., 2006; Cohn and Lapata, 2007; Zhao et al., 2008; Callison-Burch, 2008; Guzmán and Garrido, 2008). The procedure to extract paraphrases in these approaches is similar to standard phrase extraction in SMT systems, and therefore a large amount of additional parallel corpus is required. Moreover, as discussed in Section 5, when unknown texts are not from the same domain as the SMT training corpus, it is likely that paraphrases found through such methods will yield misleading translations.

Bond et al. (2008) use grammars to paraphrase the whole source sentence, covering aspects like word order and minor lexical variations (tenses etc.), but not content words. The paraphrases are added to the source side of the corpus and the corresponding target sentences are duplicated. This, however, may yield distorted probability estimates in the phrase table, since these were not computed from parallel data.

The main use of monolingual paraphrases in MT to date has been for evaluation. For example, (Kauchak and Barzilay, 2006) paraphrase references to make them closer to the system translation in order to obtain more reliable results when using automatic evaluation metrics like BLEU (Papineni et al., 2002).

2.3 Textual Entailment and Entailment Rules

Textual Entailment (TE) has recently become a prominent paradigm for modeling semantic inference, capturing the needs of a broad range of text understanding applications (Giampiccolo et al., 2007). Yet, its application to SMT has been so far limited to MT evaluation (Pado et al., 2009).

TE defines a directional relation between two texts, where the meaning of the entailed text (*hypothesis*, *h*) can be inferred from the meaning of the entailing *text, t*. Under this paradigm, paraphrases are a special case of the entailment relation, when the relation is symmetric (the texts entail each other). Otherwise, we say that one text *directionally entails* the other.

A common practice for proving (or generating) *h* from *t* is to apply *entailment rules* to *t*. An entailment rule, denoted *LHS* $\Rightarrow$ *RHS*, specifies an entailment relation between two text fragments (the Left- and Right- Hand Sides), possibly with variables (e.g. *build* X *in* $Y \Rightarrow X$ *is completed in* Y). A paraphrasing rule is denoted with $\Leftrightarrow$. When a rule is applied to a text, a new text is inferred, where the matched LHS is replaced with the RHS. For example, the rule *skyscraper* $\Rightarrow$ *building* is applied to "*The world's tallest skyscraper was completed in Taiwan*" to infer "*The world's tallest building was completed in Taiwan*". In this work, we employ lexical entailment rules, i.e. rules without variables. Various resources for lexical rules are available, and the prominent one is WordNet (Fellbaum, 1998), which has been used in virtually all TE systems (Giampiccolo et al., 2007).

Typically, a rule application is valid only under specific contexts. For example, *mouse* $\Rightarrow$ *rodent* should not be applied to "*Use the mouse to mark your answers*". Context-models can be exploited to validate the application of a rule to a text. In such models, an explicit Word Sense Disambiguation (WSD) is not necessarily required; rather, an implicit sense-match is sought after (Dagan et al., 2006). Within the scope of our paper, rule application is handled similarly to Lexical Substitution (McCarthy and Navigli, 2007), considering the contextual relationship between the text and the rule. However, in general, entailment rule application addresses other aspects of context matching as well (Szpektor et al., 2008).

3 Textual Entailment for Statistical Machine Translation

Previous solutions for handling unknown terms in a source text s augment the SMT system's phrase table based on multilingual corpora. This allows indirectly paraphrasing s, when the SMT system chooses to use a paraphrase included in the table and produces a translation with the corresponding target phrase for the unknown term.

We propose using monolingual paraphrasing methods and resources for this task to obtain a more extensive set of rules for paraphrasing the source. These rules are then applied to s directly to produce alternative versions of the source text prior to the translation step. Moreover, further coverage increase can be achieved by employing directional entailment rules, when paraphrasing is not possible, to generate more general texts for translation.

Our approach, based on the textual entailment framework, considers the newly generated texts as entailed from the original one. Monolingual semantic resources such as WordNet can provide entailment rules required for both these symmetric and asymmetric entailment relations.

Input: A text t with one or more unknown terms;
a monolingual resource of entailment rules;
k - maximal number of source alternatives to produce

Output: A translation of either (in order of preference):
a paraphrase of t OR a text entailed by t OR t itself

1. For each unknown term - fetch entailment rules:
   (a) Fetch rules for paraphrasing; disregard rules whose RHS is not in the phrase table
   (b) If the set of rules is empty: fetch directional entailment rules; disregard rules whose RHS is not in the phrase table
2. Apply a context-model to compute a score for each rule application
3. Compute total source score for each entailed text as a combination of individual rule scores
4. Generate and translate the top-k entailed texts
5. If $k > 1$
   (a) Apply target model to score the translation
   (b) Compute final source-target score
6. Pick highest scoring translation

Figure 1: Scheme for handling unknown terms by using monolingual resources through textual entailment

Through the process of applying entailment rules to the source text, multiple alternatives of entailed texts are generated. To rank the candidate texts we employ monolingual context-models to provide scores for rule applications over the source sentence. This can be used to (a) directly select the text with the highest score, which can then be translated, or (b) to select a subset of top candidates to be translated, which will then be ranked using the target language information as well. This pruning reduces the load of the SMT system, and allows for potential improvements in translation quality by considering both source- and target-language information.

The general scheme through which we achieve these goals, which can be implemented using different context-models and scoring techniques, is detailed in Figure 1. Details of our concrete implementation are given in Section 4.

Preliminary analysis confirmed (as expected) that readers prefer translations of paraphrases, when available, over translations of directional entailments. This consideration is therefore taken into account in the proposed method.

The input is a text unit to be translated, such as a sentence or paragraph, with one or more unknown terms. For each unknown term we first fetch a list of candidate rules for paraphrasing (e.g. synonyms), where the unknown term is the LHS. For example, if our unknown term is *dodge*, a possible candidate might be *dodge* $\Leftrightarrow$ *circumvent*. We inflect the RHS to keep the original morphological information of the unknown term and filter out rules where the inflected RHS does not appear in the phrase table (step 1a in Figure 1).

When no applicable rules for paraphrasing are available (1b), we fetch directional entailment rules (e.g. hypernymy rules such as *dodge* $\Rightarrow$ *avoid*), and filter them in the same way as for paraphrasing rules. To each set of rules for a given unknown term we add the "identity-rule", to allow leaving the unknown term unchanged, the correct choice in cases of proper names, for example.

Next, we apply a context-model to compute an applicability score of each rule to the source text (step 2). An entailed text's total score is the combination (e.g. product, see Section 4) of the scores of the rules used to produce it (3). A set of the top-k entailed texts is then generated and sent for translation (4).

If more than one alternative is produced by the source model (and $k > 1$), a target model is applied on the selected set of translated texts (5a). The combined source-target model score is a combination of the scores of the source and target models (5b). The final translation is selected to be the one that yields the highest combined source-target score (6). Note that setting $k = 1$ is equivalent to using the source-language model alone.

Our algorithm validates the application of the entailment rules at two stages – before and after translation, through context-models applied at each end. As the experiments will show in Section 4, a large number of possible combinations of entailment rules is a common scenario, and therefore using the source context models to reduce this number plays an important role.

4 Experimental Setting

To assess our approach, we conducted a series of experiments; in each experiment we applied the scheme described in 3, changing only the models being used for scoring the generated and translated texts. The setting of these experiments is described in what follows.

SMT data To produce sentences for our experiments, we use Matrax (Simard et al., 2005), a standard phrase-based SMT system, with the exception that it allows gaps in phrases. We use approximately 1M sentence pairs from the English-French

Europarl corpus for training, and then translate a test set of 5,859 English sentences from the News corpus into French. Both resources are taken from the shared translation task in WMT-2008 (Callison-Burch et al., 2008). Hence, we compare our method in a setting where the training and test data are from different domains, a common scenario in the practical use of MT systems.

Of the 5,859 translated sentences, 2,494 contain unknown terms (considering only sequences with alphabetic symbols), summing up to 4,255 occurrences of unknown terms. 39% of the 2,494 sentences contain more than a single unknown term.

Entailment resource We use WordNet 3.0 as a resource for entailment rules. Paraphrases are generated using synonyms. Directionally entailed texts are created using hypernyms, which typically conform with entailment. We do not rely on sense information in WordNet. Hence, any other semantic resource for entailment rules can be utilized.

Each sentence is tagged using the OpenNLP POS tagger[2]. Entailment rules are applied for unknown terms tagged as nouns, verbs, adjectives and adverbs. The use of relations from WordNet results in 1,071 sentences with applicable rules (with phrase table entries) for the unknown terms when using synonyms, and 1,643 when using both synonyms and hypernyms, accounting for 43% and 66% of the test sentences, respectively.

The number of alternative sentences generated for each source text varies from 1 to 960 when paraphrasing rules were applied, and reaches very large numbers, up to 89,700 at the "worst case", when all TE rules are employed, an average of 456 alternatives per sentence.

Scoring source texts We test our proposed method using several context-models shown to perform reasonably well in previous work:

- ***FREQ***: The first model we use is a context-independent baseline. A common useful heuristic to pick an entailment rule is to select the candidate with the highest frequency in the corpus (Mccarthy et al., 2004). In this model, a rule's score is the normalized number of occurrences of its RHS in the training corpus, ignoring the context of the LHS.
- ***LSA***: Latent Semantic Analysis (Deerwester et al., 1990) is a well-known method for representing the contextual usage of words based on corpus statistics. We represented each term by a normalized vector of the top 100 SVD dimensions, as described in (Gliozzo, 2005). This model measures the similarity between the sentence words and the RHS in the LSA space.
- ***NB***: We implemented the unsupervised Naïve Bayes model described in (Glickman et al., 2006) to estimate the probability that the unknown term entails the RHS in the given context. The estimation is based on corpus co-occurrence statistics of the context words with the RHS.
- ***LMS***: This model generates the Language Model probability of the RHS in the source. We use 3-grams probabilities as produced by the SRILM toolkit (Stolcke, 2002).

Finally, as a simple baseline, we generated a random score for each rule application, ***RAND***.

The score of each rule application by any of the above models is normalized to the range (0,1]. To combine individual rule applications in a given sentence, we use the product of their scores. The monolingual data used for the models above is the source side of the training parallel corpus.

Target-language scores On the target side we used either a standard 3-gram language-model, denoted ***LMT***, or the score assigned by the complete SMT log-linear model, which includes the language model as one of its components (***SMT***).

A pair of a *source*:*target* models comprises a complete model for selecting the best translated sentence, where the overall score is the product of the scores of the two models.

We also applied several combinations of source models, such as *LSA* combined with *LMS*, to take advantage of their complementary strengths. Additionally, we assessed our method with source-only models, by setting the number of sentences to be selected by the source model to one ($k = 1$).

5 Results

5.1 Manual Evaluation

To evaluate the translations produced using the various source and target models and the different rule-sets, we rely mostly on manual assessment, since automatic MT evaluation metrics like BLEU do not capture well the type of semantic variations

[2] http://opennlp.sourceforge.net

	Model	Precision (%)		Coverage (%)	
		PARAPH.	TE	PARAPH.	TE
1	*–:SMT*	75.8	73.1	32.5	48.1
2	*NB:SMT*	75.2	71.5	32.3	47.1
3	*LSA:SMT*	74.9	72.4	32.1	47.7
4	*NB:–*	74.7	71.1	32.1	46.8
5	*LMS:LMT*	73.8	70.2	31.7	46.3
6	*FREQ:–*	72.5	68.0	31.2	44.8
7	*RAND*	57.2	63.4	24.6	41.8

Table 1: Translation acceptance when using only paraphrases and when using all entailment rules. "*:*" indicates which model is applied to the source (left side) and which to the target language (right side).

generated in our experiments, particularly at the sentence level.

In the manual evaluation, two native speakers of the target language judged whether each translation preserves the meaning of its reference sentence, marking it as acceptable or unacceptable. From the sentences for which rules were applicable, we randomly selected a sample of sentences for each annotator, allowing for some overlapping for agreement analysis. In total, the translations of 1,014 unique source sentences were manually annotated, of which 453 were produced using only hypernyms (no paraphrases were applicable). When a sentence was annotated by both annotators, one annotation was picked randomly.

Inter-annotator agreement was measured by the percentage of sentences the annotators agreed on, as well as via the Kappa measure (Cohen, 1960). For different models, the agreement rate varied from 67% to 78% (72% overall), and the Kappa value ranged from 0.34 to 0.55, which is comparable to figures reported for other standard SMT evaluation metrics (Callison-Burch et al., 2008).

Translation with TE For each model m, we measured $Precision_m$, the percentage of acceptable translations out of all sampled translations. $Precision_m$ was measured both when using only paraphrases (PARAPH.) and when using all entailment rules (TE). We also measured $Coverage_m$, the percentage of sentences with acceptable translations, A_m, out of all sentences (2,494). As our annotators evaluated only a sample of sentences, A_m is estimated as the model's total number of sentences with applicable rules, S_m, multiplied by the model's Precision (S_m was 1,071 for paraphrases and 1,643 for entailment rules): $Coverage_m = \frac{S_m \cdot Precision_m}{2{,}494}$.

Table 1 presents the results of several source-target combinations when using only paraphrases and when also using directional entailment rules. When all rules are used, a substantial improvement in coverage is consistently obtained across all models, reaching a relative increase of 50% over paraphrases only, while just a slight decrease in precision is observed (see Section 5.3 for some error analysis). This confirms our hypothesis that directional entailment rules can be very useful for replacing unknown terms.

For the combination of source-target models, the value of k is set depending on which rule-set is used. Preliminary analysis showed that $k = 5$ is sufficient when only paraphrases are used and $k = 20$ when directional entailment rules are also considered.

We measured statistical significance between different models for precision of the TE results according to the Wilcoxon signed ranks test (Wilcoxon, 1945). Models 1-6 in Table 1 are significantly better than the *RAND* baseline ($p < 0.03$), and models 1-3 are significantly better than model 6 ($p < 0.05$). The difference between *–:SMT* and *NB:SMT* or *LSA:SMT* is not statistically significant.

The results in Table 1 therefore suggest that taking a source model into account preserves the quality of translation. Furthermore, the quality is maintained even when source models' selections are restricted to a rather small top-k ranks, at a lower computational cost (for the models combining source and target, like *NB:SMT* or *LSA:SMT*). This is particularly relevant for on-demand MT systems, where time is an issue. For such systems, using this source-language based pruning methodology will yield significant performance gains as compared to target-only models.

We also evaluated the baseline strategy where unknown terms are omitted from the translation, resulting in 25% precision. Leaving unknown words untranslated also yielded very poor translation quality in an analysis performed on a similar dataset.

Comparison to related work We compared our algorithm with an implementation of the algorithm proposed by (Callison-Burch et al., 2006) (see Section 2.2), henceforth *CB*, using the Spanish side of Europarl as the pivot language.

Out of the tested 2,494 sentences with unknown terms, *CB* found paraphrases for 706 sentences (28.3%), while with any of our models, including

Model	Precision (%)	Coverage (%)	Better (%)
NB:SMT (TE)	85.3	56.2	72.7
CB	85.3	24.2	12.7

Table 2: Comparison between our top model and the method by Callison-Burch et al. (2006), showing the percentage of times translations were considered acceptable, the model's coverage and the percentage of times each model scored better than the other (in the 14% remaining cases, both models produced unacceptable translations).

Model	BLEU (TE)	METEOR (TE)
–:SMT	15.50	0.1325
NB:SMT	15.37	0.1316
LSA:SMT	15.51	0.1318
NB:–	15.37	0.1311
CB	15.33	0.1299
Baseline SMT	15.29	0.1294

Table 3: Performance of the best models according to automatic MT evaluation metrics at the corpus level. The baseline refers to translation of the text without applying any entailment rules.

NB:SMT, our algorithm found applicable entailment rules for 1,643 sentences (66%).

The quality of the *CB* translations was manually assessed for a sample of 150 sentences. Table 2 presents the precision and coverage on this sample for both *CB* and *NB:SMT*, as well as the number of times each model's translation was preferred by the annotators. While both models achieve equally high precision scores on this sample, the *NB:SMT* model's translations were undoubtedly preferred by the annotators, with a considerably higher coverage.

With the *CB* method, given that many of the phrases added to the phrase table are noisy, the global quality of the sentences seem to have been affected, explaining why the judges preferred the *NB:SMT* translations. One reason for the lower coverage of *CB* is the fact that paraphrases were acquired from a corpus whose domain is different from that of the test sentences. The entailment rules in our models are not limited to paraphrases and are derived from WordNet, which has broader applicability. Hence, utilizing monolingual resources has proven beneficial for the task.

5.2 Automatic MT Evaluation

Although automatic MT evaluation metrics are less appropriate for capturing the variations generated by our method, to ensure that there was no degradation in the system-level scores according to such metrics we also measured the models' performance using BLEU and METEOR (Agarwal and Lavie, 2007). The version of METEOR we used on the target language (French) considers the stems of the words, instead of surface forms only, but does not make use of WordNet synonyms.

We evaluated the performance of the top models of Table 1, as well as of a baseline SMT system that left unknown terms untranslated, on the sample of 1,014 manually annotated sentences. As shown in Table 3, all models resulted in improvement with respect to the original sentences (baseline). The difference in METEOR scores is statistically significant ($p < 0.05$) for the three top models against the baseline. The generally low scores may be attributed to the fact that training and test sentences are from different domains.

5.3 Discussion

The use of entailed texts produced using our approach clearly improves the quality of translations, as compared to leaving unknown terms untranslated or omitting them altogether. While it is clear that textual entailment is useful for increasing coverage in translation, further research is required to identify the amount of *information loss* incurred when non-symmetric entailment relations are being used, and thus to identify the cases where such relations are detrimental to translation.

Consider, for example, the sentence: "*Conventional military models are geared to* ***decapitate*** *something that, in this case, has no head.*". In this sentence, the unknown term was replaced by *kill*, which results in missing the point originally conveyed in the text. Accordingly, the produced translation does not preserve the meaning of the source, and was considered unacceptable: "*Les modèles militaires visent à* ***faire*** *quelque chose que, dans ce cas, n'est pas responsable.*".

In other cases, the selected hypernyms were too generic words, such as *entity* or *attribute*, which also fail to preserve the sentence's meaning. On the other hand, when the unknown term was a very specific word, hypernyms played an important role. For example, "*Bulgaria is the most sought-after east European real estate target, with its low-cost ski* ***chalets*** *and oceanfront homes*". Here, *chalets* are replaced by *houses* or *units* (depending on the model), providing a translation that would be acceptable by most readers.

Other incorrect translations occurred when the unknown term was part of a phrase, for example, *troughs* replaced with *depressions* in *peaks*

and troughs, a problem that also strongly affects paraphrasing. In another case, *movement* was the hypernym chosen to replace *labor* in *labor movement*, yielding an awkward text for translation.

Many of the cases which involved ambiguity were resolved by the applied context-models, and can be further addressed, together with the above mentioned problems, with better source-language context models.

We suggest that other types of entailment rules could be useful for the task beyond the straightforward generalization using hypernyms, which was demonstrated in this work. This includes other types of lexical entailment relations, such as holonymy (e.g. *Singapore* $\Rightarrow$ *Southeast Asia*) as well as lexical syntactic rules (*X cure Y* $\Rightarrow$ *treat Y with X*). Even syntactic rules, such as clause removal, can be recruited for the task: *"Obama, the 44th president, declared Monday . . . "* $\Rightarrow$ *"Obama declared Monday . . . "*. When the system is unable to translate a term found in the embedded clause, the translation of the less informative sentence may still be acceptable by readers.

6 Conclusions and Future Work

In this paper we propose a new entailment-based approach for addressing the problem of unknown terms in machine translation. Applying this approach with lexical entailment rules from WordNet, we show that using monolingual resources and textual entailment relationships allows substantially increasing the quality of translations produced by an SMT system. Our experiments also show that it is possible to perform the process efficiently by relying on source language context-models as a filter prior to translation. This pipeline maintains translation quality, as assessed by both human annotators and standard automatic measures.

For future work we suggest generating entailed texts with a more extensive set of rules, in particular lexical-syntactic ones. Combining rules from monolingual and bilingual resources seems appealing as well. Developing better context-models to be applied on the source is expected to further improve our method's performance. Specifically, we suggest taking into account the prior likelihood that a rule is correct as part of the model score.

Finally, some researchers have advocated recently the use of shared structures such as parse forests (Mi and Huang, 2008) or word lattices (Dyer et al., 2008) in order to allow a compact representation of alternative inputs to an SMT system. This is an approach that we intend to explore in future work, as a way to efficiently handle the different source language alternatives generated by entailment rules. However, since most current MT systems do not accept such type of inputs, we consider the results on pruning by source-side context models as broadly relevant.

Acknowledgments

This work was supported in part by the ICT Programme of the European Community, under the PASCAL 2 Network of Excellence, ICT-216886 and The Israel Science Foundation (grant No. 1112/08). We wish to thank Roy Bar-Haim and the anonymous reviewers of this paper for their useful feedback. This publication only reflects the authors' views.

References

Abhaya Agarwal and Alon Lavie. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of WMT-08*.

Karunesh Arora, Michael Paul, and Eiichiro Sumita. 2008. Translation of Unknown Words in Phrase-Based Statistical Machine Translation for Languages of Rich Morphology. In *Proceedings of SLTU*.

Francis Bond, Eric Nichols, Darren Scott Appling, and Michael Paul. 2008. Improving Statistical Machine Translation by Paraphrasing the Training Data. In *Proceedings of IWSLT*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of HLT-NAACL*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of WMT*.

Chris Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In *Proceedings of EMNLP*.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of ACL*.

Ido Dagan, Oren Glickman, Alfio Massimiliano Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. 2006. Direct Word Sense Matching for Lexical Substitution. In *Proceedings of ACL*.

Scott Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.A. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing Word Lattice Translation. In *Proceedings of ACL-HLT*.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2008. Communicating Unknown Words in Machine Translation. In *Proceedings of LREC*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognising Textual Entailment Challenge. In *Proceedings of ACL-WTEP Workshop*.

Oren Glickman, Ido Dagan, Mikaela Keller, Samy Bengio, and Walter Daelemans. 2006. Investigating Lexical Substitution Scoring for Subtitle Generation. In *Proceedings of CoNLL*.

Alfio Massimiliano Gliozzo. 2005. *Semantic Domains in Computational Linguistics*. Ph.D. thesis, University of Trento.

Francisco Guzmán and Leonardo Garrido. 2008. Translation Paraphrases in Phrase-Based Machine Translation. In *Proceedings of CICLing*.

Nizar Habash. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. In *Proceedings of ACL-HLT*.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of HLT-NAACL*.

Kevin Knight and Jonathan Graehl. 1997. Machine Transliteration. In *Proceedings of ACL*.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of EACL*.

Philippe Langlais and Alexandre Patry. 2007. Translating Unknown Words by Analogical Learning. In *Proceedings of EMNLP-CoNLL*.

Dekang Lin and Patrick Pantel. 2001. DIRT – Discovery of Inference Rules from Text. In *Proceedings of SIGKDD*.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of SemEval*.

Diana Mccarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of ACL*.

Haitao Mi and Liang Huang. 2008. Forest-based Translation Rule Extraction. In *Proceedings of EMNLP*.

Sebastian Pado, Michel Galley, Daniel Jurafsky, and Christopher D. Manning. 2009. Textual Entailment Features for Machine Translation Evaluation. In *Proceedings of WMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*.

M. Simard, N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, and K. Yamada. 2005. Translating with Non-contiguous Phrases. In *Proceedings of HLT-EMNLP*.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP*.

Idan Szpektor, Ido Dagan, Roy Bar-Haim, and Jacob Goldberger. 2008. Contextual Preferences. In *Proceedings of ACL-HLT*.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages. In *Proceedings of EACL*.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008. Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. In *Proceedings of ACL-HLT*.