Ambiguity Reduction for Machine Translation: Human-Computer Collaboration

Marcus Sammer[†] **Kobi Reiter**[†] **Stephen Soderland**[†] **Katrin Kirchhoff**[‡] **Oren Etzioni**[†]

Turing Center[†]
University of Washington
Seattle, WA 98195

Electrical Engineering[‡]
University of Washington
Seattle, WA 98195

{sammer,dbgalur,soderlan,katrin,etzioni}@{cs[†]/ee[‡]}.washington.edu

Abstract

Statistical Machine Translation (SMT) accuracy degrades when there is only a limited amount of training, or when the training is not from the same domain or genre of text as the target application. However, cross-domain applications are typical of many real world tasks. We demonstrate that SMT accuracy can be improved in a cross-domain application by using a controlled language (CL) interface to help reduce lexical ambiguity in the input text. Our system, CL-MT, presents a monolingual user with a choice of word senses for each content word in the input text. CL-MT temporarily adjusts the underlying SMT system's phrase table, boosting the scores of translations that include the word senses preferred by the user and lowering scores for disfavored translations. We demonstrate that this improves translation adequacy in 33.8% of the sentences in Spanish to English translation of news stories, where the SMT system was trained on proceedings of the European Parliament.

1 Introduction

Statistical Machine Translation (SMT) is sensitive to the genre and domain of the training data that is used to train the translation model. The best performance is typically achieved when the texts to be translated are drawn from the same population of texts as the training data. Unfortunately, many real world applications are for target domains or genres for which readily available parallel training corpora do not exist. Mismatches between training and test data result in deteriorated performance.

One source of translation errors is lexical ambiguity in the input text, which may result in lexical errors in the translation. State-of-the-art SMT systems use a phrase-based approach to translation (Och et al., 1999; Koehn et al., 2003; Tillmann, 2003; Zens and Ney, 2004), where translations are obtained by concatenating translations of chunks of words (as opposed to single words) in the input sentence. When training and test data are matched, a phrase-based SMT system can implicitly perform word-sense disambiguation (WSD) and choose the correct translation because the local context of the input word is taken into consideration. In fact, additional explicit WSD has not been shown to be helpful (e.g. Carpuat and Wu, 2005). Under mismatched conditions, however, lexical ambiguity may become a much more significant source of errors because word senses occurring in the test data may never have occurred in similar context in the training data.

In this paper we present CL-MT, a hybrid controlled language-statistical MT system where cross-domain SMT is improved by human guided WSD. This study is part of a larger research effort on utilizing machine translation technology to enhance human-human communication. A typical application scenario is on-the-fly automatic email translation, where two users that do not share the same language are engaged in an email exchange,

and an automatic MT system is used to translate typed input while the email message is being composed. Such applications need to handle data from a wide range of domains, and we cannot assume that in-domain data will be available in all cases. Thus, cross-domain application of MT components will be the norm. On the other hand, we can assume that a monolingual human user will not only be available but will also be motivated to assist in improving the automatic translation.

Our proposed CL-MT system generates a controlled language (CL) lexicon where each entry is for a distinct word sense of a term in the source language, and entries are associated with one or more possible translations of that word sense into the target language. CL-MT has a graphical user interface that presents a monolingual speaker with alternate senses for words in an input sentence. Scores in an SMT phrase table are then boosted for translations associated with the selected word senses and scores are decreased for those word senses that were not selected.

One key feature of CL-MT is that it is both domain-independent and genre-independent. This distinguishes it from other controlled language systems that operate in a narrow domain, typically on technical specifications (Nyberg and Mitamura, 1996; Fuchs et al., 1998; Schwitter, 2002). In a narrow domain, a CL lexicon can assume that nearly all words have a single word sense – open domain CL must handle ambiguity in nearly all words.

In this paper we demonstrate that:

1. Human lexical disambiguation can improve translation accuracy in an SMT system.
2. A CL lexicon can be created automatically that gives useful, intuitive word sense choices to a monolingual user.
3. This can be done in a domain-independent, genre-independent manner.

Our empirical evaluation of cross-domain translation from Spanish to English shows that using the CL-MT system yielded a statistically significant improvement in translation adequacy. Adequacy improved for 33.8% of the sentences over a baseline SMT system, while it was reduced for 12.1% of the sentences. Section 2 describes creating the CL lexicon. This is followed by an explanation of how the user's preferences are incorporated into the underlying SMT system in Section 3. Empirical results are presented in Section 4. The paper finishes with Sections 5 and 6 discussing related research and future work respectively.

2 Creating a CL Lexicon

The first step in adapting CL-MT to a new language pair is to create a CL lexicon. This serves as the basis for the alternate word senses that CL-MT's interface presents to the user. The desired characteristics of a CL lexicon are the following:

- Each entry is a distinct word sense for a word in the source language (L1).
- Each entry is associated with one or more translations of that word into the target language (L2).
- Each entry has a short, intuitive gloss or set of cue words that enables a user to select the entry with the intended word sense.

We experimented with various methods of creating such a CL lexicon. The method that gave best results requires a bilingual dictionary and a machine readable dictionary (MRD) for L2 that provides glosses that are then translated into L1. We continue to work on methods that do not require an MRD, and thus scale to a larger set of languages.

2.1 Lexicon with MRD Glosses

CL-MT begins by looking up each source word *s* in a bilingual dictionary. We used a dictionary from UltraLingua for our Spanish-English experiments (http://www.ultralingua.net). The system then looks up each translation *t* in an MRD for language L2. This gives one or more distinct word senses for *t*, each with a gloss written by a lexicographer. The system translates these glosses from L2 into L1.

Each entry in the CL lexicon is for a distinct word sense, where each entry has a meaningful gloss in L1 (which, of course, may be poorly translated). The algorithm for creating a CL lexicon with MRD glosses is shown in Figure 1. We used WordNet 2.0 (wordnet.princeton.edu/w3wn.html) as our MRD for experiments in which the target language was English. WordNet has good cover-

```
For each term s in the source language {
    Look up s in a bilingual dictionary
    For each translation t of s {
        Look up t in a target language MRD
        For each of the first k glosses g of t {
            Translate g into the source language
            Create a CL lexicon entry with s, g, t
            Merge entries with matching s, g
        }
    }
}
```

Figure 1. Algorithm to create a CL lexicon using a bilingual dictionary and a machine readable dictionary for the target language.

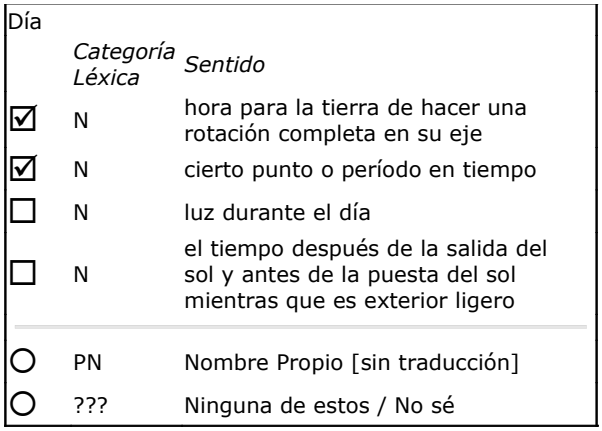

Figure 2. CL lexicon entry for the Spanish word "día". The first two entries correspond to the English translation "day", the third entry corresponds to "daylight", and the fourth entry corresponds to "day", "daylight" or "daytime"

age, although only for nouns, verbs, adjectives, and adverbs. It has a separate entry for each word sense, with the more common senses listed first, and rare usages towards the end of the list. We created CL lexicon entries for the first k WordNet entries for each part of speech for each translation *t*, with k set to 3 if there were more than 5 word senses and k set to 2 otherwise. WordNet glosses are often extremely long, so we truncated the gloss at the first semicolon.

An example entry for a CL lexicon for Spanish to English is shown in Figure 2. This is for the Spanish word "día", which has three translations according to the bilingual dictionary: "day", "daylight", and "daytime".

The first two entries ("time for Earth to make a complete rotation on its axis" and "some point or period in time") have the English translation "day". The next entry ("light during the daytime") is for the English "daylight". The last entry ("the time after sunrise and before sunset …") has three translations, "day", "daylight" and "daytime".

We used Google Translator to translate glosses into L1 (http://www.google.com/language_tools). This could also have been done with the SMT system that was trained for our CL-MT system.

Since the L2 MRD gives all senses for a word in L2, some of the senses may be inappropriate for their corresponding L1 translations. The CL lexicon will contain these inappropriate word senses, but because of the clear glosses provided by the MRD, a monolingual source language speaker is easily able to disregard the incorrect senses.

Because WordNet only contains entries for nouns, verbs, adjectives, and adverbs, we also provided hand written entries for a few dozen pronouns that specify gender and number. The CL interface also gives the user two additional options for each word: to leave it untranslated as a proper noun, or to leave the word unannotated.

The current implementation of CL-MT includes only single words in its lexicon. This is problematic for phrases whose meanings are not compositional. In such cases, none of the word senses of the individual words in a phrase are appropriate. Future work is needed to extend the CL lexicon to handle multi-word phrases.

2.2 Lexicon with Back Translation Cues

We also experimented with a lexicon building method that does not assume an MRD for the target language. CL-MT begins by looking up each source word *s* in a bilingual dictionary, and creating a separate entry for each translation *t* of *s*. In place of a gloss, each entry has two sets of cue words: 1) back translations of *t* into the source language and 2) context words of *s* that are translations of context words of *t*.

Figure 3 shows the Spanish source word "día", where the first entry corresponds to the English translation "day" which has back translations into Spanish of "día" and "fecha" (date). The second entry corresponds to "daylight", and has back

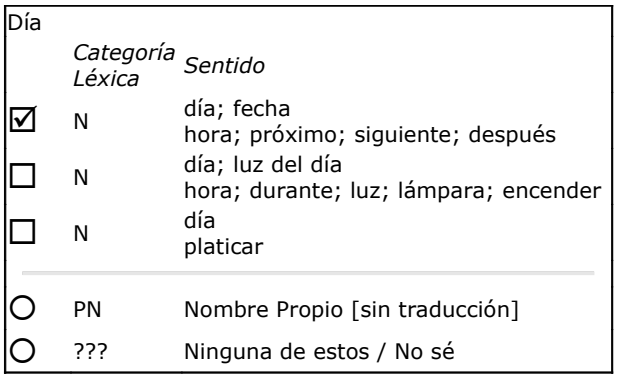

Figure 3. An entry for the Spanish word "día" using back translations and context words as cues.

translations of "día" and "luz del día" (light of day). The third entry corresponds to "daytime", which has no back translations other than the original word "día".

We had mixed results with this method for many of the entries in the CL lexicon. For some words and some lexicon entries, either the back translations or the context cues provided clear information to the user. For other entries, this was not the case.

Another difficulty with the back translation method is that the L2 translations themselves are often ambiguous. This leads to lists of back translations that include extraneous meanings along with the intended word sense. As opposed to the MRD method, we have no simple method for separating out the inappropriate senses so they can be disregarded with out also throwing out the correct sense. For example, the Spanish word "enlace" has the intended meaning of "link" to a Web page. Unfortunately, the back translations of link include "campo de golf" (golf course), which may lead a user to reject this as the wrong sense of the word "enlace".

3 Influencing the SMT Decoding

The final step is to use the output of the CL interface to bias the SMT system to favor translations that reflect the word sense intended by the user.

3.1 Baseline SMT System

For our experiments we used a phrase-based SMT system (Kirchhoff and Yang, 2005) based on the public-domain decoder, Pharaoh (Koehn, 2004), that utilizes a log-linear combination of feature scores. The translation model was trained on 15M words of parallel Spanish-English European Parliament proceedings. The model combines two lexical and two phrasal translation scores (one for each translation direction), a phrase length penalty, a word transition penalty, a distortion score and a language model score. Score combination weights were optimized on a development set from the parliamentary proceedings domain. The language model was trained on the English side of the training corpus. Thus, none of the system components were tuned to the new domain (news text). The system has a state-of-the-art performance (around 31% BLEU score) on a standard benchmark task for the Europarl corpus (Koehn and Monz, 2005). For the present experiments, single-pass monotone decoding was used. This disallows word reordering and ignores potential benefits from more advances models (i.e. higher-order language models) but results in faster decoding, which may be crucial for real-world applications. On the out-of-domain test set used in the experiments described below, the system obtained a BLEU score of 21.7%, with a 95% confidence interval from 18.7% to 24.7% on a test set of 198 sentences.

3.2 Using Output of the CL Interface

CL-MT uses the output of the CL interface to modify the feature scores of entries in a temporary copy of the SMT phrase table. For these experiments we use the baseline system and rerun decoding with the modified phrase table. Additional parameter settings determine the degree to which the baseline feature scores are altered, depending on whether a phrase table entry includes translations from a CL lexicon that are preferred or disfavored.

If the user has annotated a word with one or more preferred senses, the *preferred translations* are those translations associated with at least one of the selected word senses. The *disfavored translations* are those where none of the word senses associated with the translation were selected by the user. The translations associated with each word sense in the lexicons are in root form, so CL-MT adds morphological variants to the lists of preferred and disfavored translations.

CL-MT creates a temporary copy of the SMT phrase table for each message. Since multiple occurrences of the same word in a message may be annotated by the user with different word senses, a unique identifier is appended to each token in the

message. Thus a word that is repeated in a message has distinct entries in the temporary phrase table for each appearance of the word in the message. For each source language phrase *sp* in the message, CL-MT looks up the target language translations of *sp* in the original phrase table. For each of these translations *tp*, CL-MT copies the *sp*-*tp* translation pair along with its corresponding feature scores into the temporary phrase table.

CL-MT modifies the temporary phrase table to ensure that words annotated as proper nouns are not translated. For each source language phrase *sp* in the temporary phrase table that contains a token *w* annotated by the user as a proper noun, if a translation *tp* of *sp* in the temporary phrase table does not contain *w*, then this *sp*-*tp* translation pair is removed from the temporary phrase table. This blocks CL-MT from translating a proper noun as something other than the source language word itself. Annotating proper nouns is necessary because our system uses an all lowercase word representation and does not contain a named entity recognition component and many components of proper names in our input language (Spanish) may be common nouns or adjectives as well.

Next, the counts shown in Figure 4 are used to modify scores in the temporary phrase table for each translation pair *sp*-*tp*. CL-MT counts the number of words in *sp* that have preferred translations, disfavored translations, or neither preferred nor disfavored translations in *tp*. This is done for for each lexicon *L*: the pronoun lexicon and the MRD generated lexicon.

pw_L = the number of tokens *w* in *sp* for which *w* is in *L* and a preferred translation of *w* appears in *tp*.

dw_L = the number of tokens *w* in *sp* for which *w* is in *L* and no preferred translation of *w* appears in *tp*, but a disfavored translation of *w* does appear in *tp*.

nw_L = the number of tokens *w* in *sp* for which *w* is in *L* and *w* is annotated with one or more preferred word senses, but none of the preferred and none of the disfavored translations of *w* appear in *tp*.

Figure 4. Counts of preferred and disfavored translations used in modifying phrase table entries.

For each lexicon *L*, the lexical and phrasal translation scores in the temporary phrase table are then multiplied by the parameter α_L for each word of *sp* in lexicon *L* that has a preferred translation in *tp*, by β_L for each word that has a disfavored translation in *tp*, and by γ_L for each word that has neither a preferred nor a disfavored translation in *tp*. More precisely, CL-MT multiplies the translation scores by

$$\prod_L \alpha_L^{pw_L} \beta_L^{dw_L} \gamma_L^{nw_L}$$

where *L* varies over the two lexicons, α_L, β_L, and γ_L are parameters, and pw_L, dw_L, and nw_L are the counts described in Figure 4.

CL-MT also handles the cases where a translation from the CL lexicon is missing from the SMT phrase table. If no preferred translation of a token *w* is found in the phrase table, CL-MT presents the preferred translations of *w* to the decoder using the XML markup facility provided by the decoder for introducing external knowledge. The decoder is allowed to bypass these suggested translations which are weighted with another parameter δ.

We tuned these parameters using a separate development set of 50 sentences. During tuning, performance was measured using the BLEU score. In addition to δ, and α_L, β_L, and γ_L for the two lexicons, CL-MT has on/off parameters for processing the proper noun annotations, for presenting translation options to the decoder using the XML markup, and for allowing the decoder to bypass these translations. Our baseline system is equivalent to setting α_L, β_L, and γ_L to 1.0, turning off the processing of proper noun annotations and turning off the presentation of translation options to the decoder using the XML markup.

4 Experimental Results

We conducted experiments to test CL-MT on Spanish sentences that were found on Web pages where there was an English translation suitable for a reference translation. These were primarily news stories, but also included press releases.

This gave us only a single reference translation, which means that in order to improve standard metrics for translation accuracy such as BLEU and position-independent word error rate (PER), precisely the same words as in the reference translations would need to be hypothesized. In order to better assess the effect of acceptable but non-

matching translation hypotheses we supplemented the automatic scores with human evaluations of the adequacy and fluency of CL-MT translations compared to translations by the baseline SMT system.

We presented human evaluators with the reference translation and two output translations: from the baseline SMT system without disambiguation, and from CL-MT. Pairs of outputs were presented in random order without indication of the system identity. The evaluators judged which output had better adequacy or if they were equal; and also judged which output had better fluency or if they were equal.

We had three fluent speakers of Spanish use CL-MT's interface to annotate a test set of 198 sentences randomly selected from our collection of Spanish news stories and press releases. Our most prolific annotator did the entire test set and the other annotators each did a portion of the same test set to help us assess inter-annotator agreement.

4.1 Parameter Tuning

We set parameters empirically for our CL-MT system as described in Section 3.2, and found values shown in Table 1 to optimize for BLEU score. The system is not sensitive to small changes in these parameter settings: even doubling a setting or reducing a setting by half makes only a small difference in performance.

From WordNet lexicon	multiply by
α_{WN} (preferred translations)	5.0
β_{WN} (disfavored translations)	0.8
γ_{WN} (neither)	1.0
From Pronoun lexicon	multiply by
α_{PR} (preferred translations)	12.0
β_{PR} (disfavored translations)	0.5
γ_{PR} (neither)	1.0
Enforce exact translation of proper nouns	
Add missing translations to phrase table with weight	0.2

Table 1. Parameter settings for CL-MT with MRD gloss. These settings strongly increase scores for preferred translations, and decrease scores for disfavored translations.

The results of these parameter settings confirm that the output of our CL interface is indeed giving useful information to an SMT system. The optimal settings are to give a strong preference to translations preferred by the user, and to avoid disfavored translations. There is also a boost in performance by knowing when not to translate a proper name. It also helps to add translations that are missing from the phrase table with a small score, although too high a score hurts performance.

4.2 Automatic Evaluation

Table 2 compares CL-MT with the baseline system. We found that CL-MT raises BLEU score and lowers PER, both indicating better translation accuracy. However, our sample was not large enough for this improvement to be statistically significant. Our CL interface is inherently labor intensive and precludes generating the large test sets common for fully automated methods.

Table 2 also shows performance of an "oracle" system. One of the authors of this paper used a version of CL-MT that displayed the English translation for each entry, and selected only entries with translations that matched words in the reference translation. If no entry matched the reference translation, the source word was left unannotated. This gives an upper bound for automatic scoring for CL-MT. We found that 3% of the content words had no CL lexicon entry due to gaps in the dictionary and that 6.6% of the entries had no word sense that was synonymous with the target translation. This was often because the meaning of words changed when used in a phrase.

Method	BLEU	PER
CL-MT	22.6	44.3
Baseline	21.7	45.0
Oracle CL-MT	27.6	39.6

Table 2. CL-MT with MRD glosses improves both BLEU score and PER. An Oracle CL-MT with perfect annotation shows the ceiling on performance gain.

4.3 Human Evaluation

We also evaluated CL-MT with human judgments, particularly since we had only one reference translation. Human evaluation takes into account synonyms and can distinguish adequacy (correct content) from fluency (correct grammar and style). Figure 5 shows two examples from the interface used for manual evaluation. At the top is the reference translation, followed by the output of CL-MT and the baseline system in random order. The evaluator does not know which method was used or who did the annotation. We used two evaluators, who had 76% agreement, and then reconciled the differences.

These examples are fairly typical of the system output. Fluency is generally comparable – the baseline is not fluent English, and CL-MT does little to improve this. When CL-MT improves adequacy, it typically improves translation of one or two words in the sentence. The translation "day of choice" is better than "agenda choice", even though both systems should have "election" rather than "choice". The second example has better adequacy from not translating the proper name, Lorenzo Rubio, and from "Spanish" rather than "Spain". Near synonyms such as "done badly" for "done wrong" do not affect adequacy.

We saw a significant improvement in adequacy from CL-MT over the baseline system, as shown in Table 3. CL-MT increased adequacy in 33.8% of the sentences, lowering adequacy in 12.1%. This means that CL-MT improved adequacy 2.8 times more often than hurting it. Some cases where human annotation hurt adequacy were from words whose meaning in a phrase was not the same as the word in isolation; some were from confusion about the meaning of a gloss; some were where annotating a word caused suboptimal translation of an adjacent word. There was no significant difference in fluency between CL-MT and the baseline system.

Table 3 also compares the "oracle" system with optimal annotation to the baseline. Our real CL-MT system improves adequacy over the baseline only a little less often than a perfect annotator would have. However, a perfect annotator increases adequacy 15.8 times more often than decreasing it.

grace period voters may not cancel their votes and then vote on election day at the polling stations .		
Fluency	**Adequacy**	**Sentence Text**
☑	☑	voters of the grace period cannot cancel their vote and vote on the day of choice in the casillas .
☑	☐	voters of the grace period cannot cancel their vote and vote on the agenda choice in the casillas .

when his father , lorenzo rubio , upset that his son had been suspended from school for two days , asked watts what his son had done wrong , he said she told him , " i don't want to hear it [spanish] in my building .		
Fluency	**Adequacy**	**Sentence Text**
☑	☑	when his father , lorenzo rubio - irritating because his son had been suspended of school for two days - asked watts what was done badly his son , the citing she replied . . ' i do not want to hear it (spanish) in my building .
☑	☐	when his father , lawrence hair - annoying because his son had been suspended of school for two days - asked watts what was done wrong his son , the citing she replied . . ' i do not want to hear it (spain) in my building .

Figure 5. Two examples from an evaluation tool to compare fluency and adequacy between CL-MT and a baseline SMT translation. The top line in each example is the reference translation, followed by baseline and CL-MT output in random order. Here the two outputs have equal fluency, but the first output has a more adequate translation of one or two phrases (e.g. "day" instead of "agenda").

Method	% better fluency	% better adequacy
CL-MT	10.6	33.8
Baseline	10.6	12.1
Both systems equal	78.8	54.0
Oracle CL-MT	13.1	39.9
Baseline	5.6	2.5
Both systems equal	81.3	57.6

Table 3. CL-MT gives a statistically significant boost to adequacy over the baseline SMT system. With perfect annotations (oracle), CL-MT rarely makes choices that hurt adequacy.

4.4 Results from Content Words Only

CL-MT can improve lexical translation in several ways: annotation of the preferred word senses of content words (nouns, verbs, adjectives, and adverbs); annotation of the preferred word senses of pronouns; enforcing non-translation of proper nouns; and supplying translations that are missing from the SMT phrase table.

Of these, the largest boost in BLEU score comes from annotation of pronouns and of proper nouns. The former is a peculiarity of our Spanish-English language pair. Pronouns in Spanish do not mark gender or number, while English pronouns do; some are ambiguously a pronoun or a determiner; and some pronouns are omitted in the English translation. Without the CL interface, a pronoun such as "su" is indiscriminately translated as "his", "her", "their", "your", "one's", or "its".

The most interesting part of CL-MT, however, is the annotation of content words. We ran CL-MT with all parameters set to neutral values except those for the MRD lexicon. This shows the contribution of our WordNet lexicon to translation accuracy without the other functionality of CL-MT. There was only a modest gain of 0.2% in BLEU score. The effect was more pronounced with human evaluation that takes into account synonymous translations.

Table 4 shows the percent of sentences where CL-MT with only the MRD lexicon improved fluency or adequacy with respect to the baseline SMT system. Disambiguation of content words accounts for 69% of the gain in adequacy from the full CL-MT and for 54% of the cases where CL-MT hurts adequacy. Annotating only content words helps adequacy 3.5 times more often than hurting it, an even better ratio than for the full CL-MT system. As before, there is no significant net affect on fluency.

Method	% better fluency	%better adequacy
CL-MT (content words)	6.6	23.2
Baseline	7.1	6.6
Both systems equal	86.4	70.2

Table 4. This shows the contribution of CL-MT where only nouns, verbs, adjectives, and adverbs are annotated.

5 Related Research

There have been several research papers recently on incorporating WSD into SMT. Carpuat and Wu conducted experiments using a WSD classifier for Chinese based on an ensemble of naïve Bayes, maximum entropy, AdaBoost, and Kernel PCA-based classifiers (Carpuat and Wu, 2005). These classifiers had a much richer feature set of contextual information than was available to the phrasal SMT system that Carpuat and Wu used. They found that BLEU scores declined when the WSD system was used to override the translation chosen by the SMT system.

A research group at Stanford (Vickrey et al., 2005) applied automatic WSD where the word senses of an English word were taken to be its possible French translations. Their system succeeded in finding the correct translation in a "fill in the blank" experiment, but did not find significant improvements in translation accuracy of full sentences.

The use of human-verified WSD has been explored by Translution.com (Orasan et al., 2005). Their method applies only to language pairs where both languages have EuroWordNet thesauri (www.illc.uva.nl/EuroWordNet). They use WordNet's interlingual index to link word senses in the source language with corresponding senses in the target language. They reported on techniques to prune out irrelevant word senses to avoid overburdening a user, but did not report on how the WSD affected translation accuracy.

A promising approach to building a CL lexicon without an MRD available is corpus-based cluster-

ing (Kikui, 1999). Kikui uses distributional clustering to identify the word sense of a source language word, and then tests each translation from a bilingual dictionary to find a translation whose context in the target language corpus best matches the context for that sense in the source language corpus.

The controlled language of CL-MT is qualitatively different than that of other research in controlled language. Our CL lexicon is designed to be domain independent and must deal directly with ambiguity of nearly all terms. Other CL systems have been developed for narrow domains, or at best, with a domain-independent architecture that relies on domain-specific knowledge.

The Kant system (Nyberg and Mitamura, 1996; Mitamura et al., 2003) was developed primarily for one-way translation of Caterpillar Tractor manuals from English. Nearly all of the content words are restricted to a single word sense, and multi-word noun phrases are only allowed if explicitly in the lexicon. Kant would reject "oil filter change" even though "oil filter" and "change" are both in the lexicon ("change of oil filter" is permitted). Attempto Controlled English (ACE) (Fuchs et al. 1998) and Processable English (PENG) (Schwitter 2002) are similarly designed for technical specifications in narrow domains.

6 Conclusions and Future Work

We have tested the hypothesis that human assistance in lexical ambiguity resolution by a monolingual source language speaker can improve translation accuracy of an SMT system. Adequacy improved for 33.8% of the sentences over a baseline SMT system, while it was reduced for 12.1% of the sentences. There was no significant difference in fluency. A small improvement in BLEU score, from 21.7% to 22.6% and a small reduction of position independent word error rate (PER) from 45.0% to 44.3% were not statistically significant. An oracle version of CL-MT shows the potential gain from optimal annotations: it improved adequacy on 39.9% of the sentences while only lowering adequacy for 2.5% of them, raising BLEU score to 27.6% and lowering PER to 39.6%.

Our experiments with CL-MT were designed as a proof of concept, so we did not formally measure the burden placed by our system on the user. In real world applications this aspect of the system becomes very important. We have a system under development that indicates to the user the word senses that the underlying SMT would choose absent any user disambiguation, easing some of the user's work.

Our CL-MT system demonstrates that human input can give a significant improvement to the adequacy of SMT translation. This performance boost can be realized if the CL lexicon provides entries for each word that distinguish separate word senses, are associated with one or more translations for each entry, and have an intuitive gloss for each entry. These criteria are met by a CL lexicon that we created using a bilingual dictionary and an MRD for the target language. The third criterion, intuitive glosses, was not met by a CL lexicon we built without using an MRD. Neither back translation cues nor context word cues allowed a user to select the correct word sense reliably.

Results from an oracle system show that there is room for greater improvement by CL-MT with better coverage by its bilingual dictionary, better morphological analysis, and a better way to handle phrases where the meaning is not compositional. For the latter problem, corpus-based techniques to find collocations may prove useful, as well as mining the SMT phrase table for phrasal translations to be included in the CL lexicon.

We are interested in pursuing methods that scale to language pairs without an MRD for either language. We are optimistic that CL-MT can be extended to any language pair where there is a simple bilingual dictionary and a corpus is available for each language. In the absence of an MRD, the main challenges are to identify distinct word senses automatically and to provide meaningful cues to the user to distinguish the word senses.

One key direction for the problem of mixed word senses is to use clustering algorithms on local context words to distinguish separate word senses (Yarowsky, 1995; Schütze, 1998), so that entries are not a mixture of partly correct and partly incorrect word senses. Using example sentences or phrases containing the word to be disambiguated may prove to be more useful descriptors than context words for aiding the user in disambiguation.

Acknowledgments

This research was carried out at the University of Washington's Turing Center, which is supported in part by a generous gift from the Utilika Foundation. The research was also funded in part by grant no. IIS-0308297 from the US National Science Foundation.

References

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. Proceedings of 43rd Annual Meeting of the ACL, pages 387-394.

Norbert E. Fuchs, Uta Schwertel, and Rolf Schwitter. 1998. Attempto Controlled English (ACE), Language Manual, Version 2.0, Institut fr Informatik, University of Zurich.

Genichiro Kikui. 1999. Resolving translation ambiguity using non-parallel bilingual corpora. Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language Processing.

Katrin Kirchhoff and Mei Yang. 2005. Improved language modeling for statistical machine translation. Proceedings of the 2005 ACL Workshop on Building and Using Parallel Texts.

Philipp Koehn and Christof Monz. 2005. Shared task: Statistical machine translation between European languages, Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 119-124.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Proceedings of HLT/NAACL, pages 127-133.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. Proceedings of AMTA 2004, pages 115-124.

Teruko Mitamura, Kathryn Baker, Eric Nyberg, and David Svoboda. 2003. Diagnostics for interactive controlled language checking. Proceedings of 4th Controlled Language Applications Workshop.

Eric H. Nyberg and Teruko Mitamura. 1996. Controlled language and knowledge-based machine translation: Principles and practice. Proceedings of First International Workshop on Controlled Language Applications.

Franz Joseph Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. Proceedings of EMNLP 1999.

Constantin Orasan, Ted Marshall, Robert Clark, Le An Ha, Ruslan Mitkov. 2005. Building a WSD Module within an MT system to enable interactive resolution in the user's source language. Proceedings of EAMT 2005.

Hinrich Schütze. 1998. Automatic Word sense disambiguation. Computational Linguistics (24,1)I, pages 97-123.

Rolf Schwitter. 2002. Developing a black box specification using controlled English. Computational Linguistics.

Christoph Tillmann. 2003. A projection extension algorithm for statistical machine translation. Proceedings of. EMNLP 2002 pages 1-8.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. Proceedings of. HLT/EMNLP 2005, pages 771-778.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of ACL 1995.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. Proceedings of HLT-NAACL 2004, pages 257-264.