Complexity of Spoken Versus Written Language for Machine Translation

Nicholas Ruiz
University of Trento
Fondazione Bruno Kessler
Trento, Italy
nicruiz@fbk.eu

Marcello Federico
Fondazione Bruno Kessler
Trento, Italy
federico@fbk.eu

Abstract

When machine translation researchers participate in evaluation tasks, they typically design their primary submissions using ideas that are not genre-specific. In fact, their systems look much the same from one evaluation campaign to another. In this paper, we analyze two popular genres: spoken language and written news, using publicly available corpora which stem from the popular WMT and IWSLT evaluation campaigns. We show that there is a sufficient amount of difference between the two genres that particular statistical modeling strategies should be applied to each task. We identify translation problems that are unique to each translation task and advise researchers of these phenomena to focus their efforts on the particular task.

1 Introduction

The machine translation community has consistently used the translation of news texts and news commentaries as some of its prime methods of evaluating the progression of MT research. News translation evaluation tasks have existed since the first NIST evaluations in the early 2000s, followed by the Workshop on Machine Translation (WMT) (Bojar et al., 2013).

In recent years, TED talks have attracted the interest of the MT research community for measuring progress. The International Workshop on Spoken Language Translation (IWSLT) is currently in its fifth year of hosting TED talk evaluation campaigns, with a growing number of translation languages and participants (Cettolo et al., 2013). Both the WMT and IWSLT evaluations have enjoyed strong performance results that have progressively improved year after year and are established today as the de-facto evaluation tasks for text and speech translation, respectively. In practice, the top performing MT systems use many of the same training and decoding approaches in these evaluations. But are the WMT and IWSLT translation tasks just different flavors of the same translation problem? Are the strategies used to translate written language directly applicable to the genre of spoken language – in particular, prepared speeches?

This paper investigates the question of what makes MT difficult for speech corpora as opposed to text corpora. We try to understand the differences between the genres of news texts and prepared speeches, both in qualitative and quantitative terms. The ultimate goal is to find information that could explain differences in MT system performance and the types of errors occurring often in MT systems trained on text and speech corpora.

We begin by surveying some of the aspects of language that make MT hard and how they relate to the problem of human understanding of text (Sections 2 and 3). We follow up the discussion with a detailed analysis to determine if these aspects are distinctive of IWSLT or WMT, or are shared in common (Sections 4-7). We contrast WMT News Commentary texts with TED talks due to their similarity to the lecture genre. We analyze their characteristics and compare them both on a monolingual and a bilingual perspective. In the monolingual perspective, we look at the characteristics of the source language that make it difficult to process. In the bilingual perspective, we look at the problem of transferring content and structure from English to German. We follow-up with a small MT experiment, comparing the performance of TED and WMT News Commentaries on similar training conditions in Section 8. In Section 9 we rec-

ommend the suitable evaluation task for various research aspects of MT, and we summarize our findings in Section 10.

2 Challenges in human readability

The most commonly researched area of language complexity lies in the field of psycholinguistics. Much of the research focuses on language acquisition and generation by native speakers or second language learners and focus on a single language.

From the reader's end, extralinguistic information such as prior world knowledge and familiarity with a topic provide context that helps her understand a text. A text can activate this information through a variety of linguistic devices, such as anaphoric mentions and grounding. Additionally, the reader must be able to organize the information received from the text into coherent blocks. Readable texts typically have a number of qualities that assist the reader in processing the information, such as redundancy, favoring concrete references over abstract principles, restatements of unfamiliar concepts, and syntactic structures appropriate for the reading level of the intended audience.

Graesser et al. (1994) introduce a *coherence assumption*, which claims that readers routinely attempt to construct coherent meanings and connections among text constituents unless the quality of the text is too poor. This concept forms one the core hypotheses in the constructivist theory of discourse comprehension. As a result, many complexity analysis tools attempt to detect coherence and cohesion through syntax, semantic, and discourse connectives (Graesser et al., 2004; Mitchell et al., 2010; Newbold and Gillam, 2010).

Biber (1988) and follow-up work by researchers investigate the variation in cohesion across text and speech corpora. Louwerse et al. (2004) perform a multi-dimensional analysis to identify a number of linguistic features that divide the corpora along several registers. Their results show variance between speech and writing corpora on a variety of factors, including type frequency, polysemy, pronoun density, abstract noun usage, type-token ratios for nouns, and stem overlap. These features divide the written and spoken genres into subdomains posing unique challenges in comprehension (e.g. prepared speeches versus conversational speech; news broadcasts versus legal documents).

3 Language Complexity in Statistical Machine Translation

Specia et al. (2011) outline three categories for features used in the task of MT quality estimation: *confidence* indicators derived from SMT models, *complexity* indicators that measure the difficulty of translating the source text, and *fluency* indicators that measure the grammaticality of a translation. Likewise, the difficulty of a translation task can be estimated by analyzing source complexity and target language features that indicate the capacity of a statistical system to generate fluent translations.

We attempt to focus on complexity issues that are irrespective of a particular text, speaker, or language pair and focus on issues that are relevant to the MT task. We can categorize these issues into three general areas: the lexicon, syntax, and semantics. When considering the lexicon, we can observe effects of vocabulary size, morphological variations, and both lexical and translation ambiguity as key impacts affecting the ability of the statistical models to cover the words in the language (Carpuat and Wu, 2007). On the syntax level, sentence length, structure complexity, and structural dependencies affect the decoding search space. On the semantic level, phenomena such as idiomatic expressions, figures of speech, anaphora, and elliptical expressions define intrinsic limitations of syntactic models. While we can observe nearly all of these language features on the monolingual level, many of these issues have a greater impact when transferring linguistic information in the process of translation. Between distant language pairs, the effects of these linguistic features cause a cumulative increase in the difficulty of MT.

Although discourse-based machine translation takes into account intersentential factors affecting translation quality (Carpuat, 2009; Foster et al., 2010), the majority of SMT systems treat each sentence independently, ruling out additional context.

4 Research methodology

In this paper, we compare two sources of spoken and written language: TED talk transcripts[1] and News Commentary texts[2]. Both types of texts cover a variety of topics whose content is produced by several authors. Although these types of texts correspond to different genres, they are popular representatives of spoken and written language investigated in MT, while belonging to similar domains. Both genres consist of speakers or authors with similar communication goals: namely, the mass distribution of information and ideas delivered by subject matter experts. At the same time, TED speakers have the additional objective of selling ideas through persuasive speeches. We focus

[1] http://www.ted.com/talks

[2] http://www.statmt.org/wmt09/translation-task.html

Measure	TED-EN	WMT-EN	TED-DE	WMT-DE
Word Count	2000018	2000016	1890106	2046071
Line Count	103588	82256	103588	82256
Surface forms	46001	50129	86787	95922
Stems	34417	36904	62929	66735
Words/Line	19.31	24.31	18.25	24.87
Stem/Surface	0.748	0.736	0.725	0.696

Table 1: Statistics for two million word TED and WMT News Commentary corpora samples.

on the English-German language pair, which belong to the same language family, but have marked differences in levels of inflection, morphological variation, verb ordering, and pronoun cases.

In our experiments, we sample approximately two million words from both the English TED and WMT News Commentary corpora, as well as the German translations of their sentences. Rather than randomly sampling sentences from the corpora, we sequentially read the sample to allow us preserve the underlying discourse. Sentences containing more than 80 words are excluded. We additionally subdivide the sampled corpora into blocks of 100,000 words to measure statistics on vocabulary growth rate.

We use TreeTagger (Schmid, 1994) to lemmatize and assign part-of-speech tags using the Penn Treebank (Marcus et al., 1993) and STTS (Schiller et al., 1995) tagsets for English and German, respectively. Some simple corpora statistics are provided in Table 1.

5 Word statistics

5.1 Sentence length

Since the unconstrained search space in SMT is exponential with respect to the length of the source sentence, we examine the distribution of sentence lengths between the TED and WMT corpora, as shown in Figure 1. On average, TED consists of lines containing around 19 words, while WMT averages five more words per line. Forty percent of the sentences in TED have between six and 15 words, while the majority of the sentences in WMT contain over 20 words. This suggests that TED is less susceptible to length-dependent decoding issues such as long distance reordering.

5.2 Predictability: Perplexity and new words

Perplexity measures the similarity of n-gram distributions between a training set and a test set. Source and target language n-gram distributions govern a SMT system's capacity to adequately translate a sequence of words with its phrase table and language model (LM). Likewise, the out-of-vocabulary (OOV) rate estimates the amount of

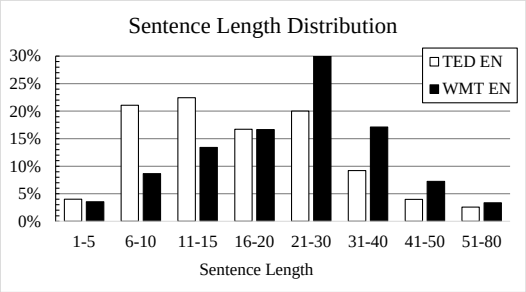


Figure 1: Sentence length statistics for English. TED talks favor shorter sentences.

source words that are impossible to translate with the given training data. We measure these notions of complexity by constructing English and German language models and evaluating their predictive power against in-domain data. Using our 2 million word corpus samples, we incrementally add 100,000 words to each corpus and evaluate its perplexity and OOV rate against a held-out 100,000 word sample from each training corpus. Using IRSTLM (Federico et al., 2008), we construct trigram LMs, using improved Kneser-Ney smoothing, no pruning, and a fixed vocabulary size of 10 million words.

According to the results shown in Figure 2, TED consistently has lower trigram perplexity rates (-46% with the full data for English, -28% for German). We observe no significant differences in OOV between TED and WMT. The results suggest TED is more capable of being modeled than WMT with the same amount of training data and the translation of TED is more regular than the translation of WMT.

6 Lexical ambiguity

Two measurements of lexical ambiguity are word polysemy and translation entropy. We analyze the ambiguity of noun and verb lemmas, which as content words carry the most important information needed to understand a sentence. We only consider

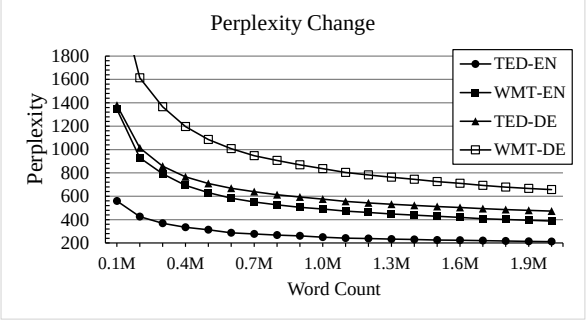


Figure 2: Perplexity change as corpus size increases for English and German.

the types that contain sense information in WordNet (Fellbaum, 1998). We take the top 100 lists of verbs and nouns from each corpus and measure their ambiguity, as described in the sections below. We compare the results against measurements on the full set of nouns or verbs and additionally measure the overlapping lemmas in the corpora.

6.1 Polysemy

As an upper-bound measure of word ambiguity, we measure the number of senses each English word in the corpus can express, as reported by WordNet. While not every sense may be observed in our corpora, this measure estimates how ambiguous a corpus is for a statistical system that considers each sense to be equally likely for a given word. Figure 3a provides a comparison between the top 100 verb and noun lemmas in the two corpora. On a global scale, we do not observe significant differences in the number of senses over the entire set of verbs and nouns in the corpora. By focusing on the top 100 lists, we observe that while the nouns and verbs shared in common between TED and WMT explain the majority of the ambiguity with respect to polysemy, the non-overlapping lemmas demonstrate TED's higher ambiguity through the use of common verbs and nouns. By isolating the lemmas that are unique to each corpus' top 100 list, we see that TED's verbs and nouns exhibit 1.5 and 2 more senses respectively than those of WMT.

In order to measure the overall effects of polysemy on the corpora, we weight the noun and verb senses by their corpora frequencies. Figure 3b shows how the distributional frequency of noun and verb senses varies over TED and WMT. For verbs, we observe that TED exhibits fewer tokens with low ambiguity and a significant increase in tokens with over 11 word senses. The noun senses behave in a similar manner, though the differences are not as pronounced.

These results demonstrate that TED favors the use of common, expressive verbs. Examples are shown in Table 2. Piantadosi et al. (2012) explain this phenomena as a trade-off between the pressures of clarity and ease in communication. We find that this is the case when combining these observations with the perplexity measures in Section 5.2.

Lemma	# Senses	TED	WMT
tell	8	2159	362
learn	6	1102	336
hear	5	875	187
read	11	529	110

Table 2: Common polysemic verbs and their occurrence frequencies in TED and WMT.

6.2 Lexical translation entropy

If the results in Section 6.1 suggest that TED talks are more ambiguous through the use of common verbs and nouns, does this transfer to the problem of SMT? To address this question, we analyze the lexical translation table provided by Moses and MGIZA through the word alignment process. We again compare TED and WMT both on the top 100 lists and the full sets of noun and verb lemmas. We train a word alignment model using MGIZA on the lemmatized corpora to build an English-German lexical translation table. In order to control the effects of alignment noise, we find the German lexical translations of each English lemma that cover the top 95% of the probability mass. Figure 4 compares TED and WMT in terms of lexical entropy.

Translating the top 100 verbs is much less ambiguous in the TED talk translation task (3.2 bits versus 3.9 bits). Most of the entropy is explained by the set of verbs TED and WMT share in common. WMT suffers from underspecification of these primarily common verbs. For example, the verb "bring", which occurs over 800 times in both corpora, exhibits an entropy of 4.04 bits and 170 translation options in TED, as opposed to 4.39 bits and 210 translation options for WMT. In terms of translation perplexity, the translation difficulty is as hard as deciding between 16 equally likely translations in TED, versus 21 in WMT. As a word with 11 senses in WordNet, this implies that fewer senses are actually being considered during translation in TED. A similar behavior can be observed for the common nouns. These results indicate that while TED has potentially higher English noun and verb polysemy, the common nouns and verbs are used more regularly than in WMT.

6.3 Pronominal anaphora

Hardmeier and Federico (2010) demonstrate that differences in the pronominal systems of a source

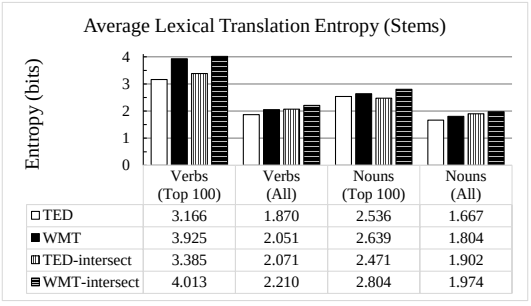
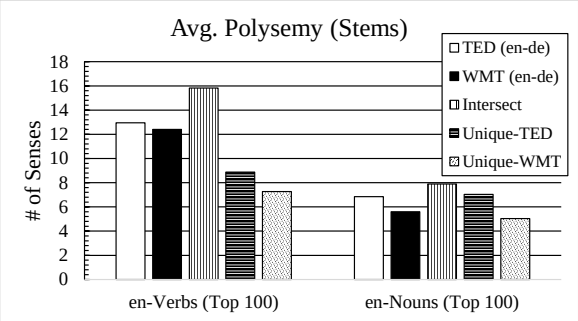
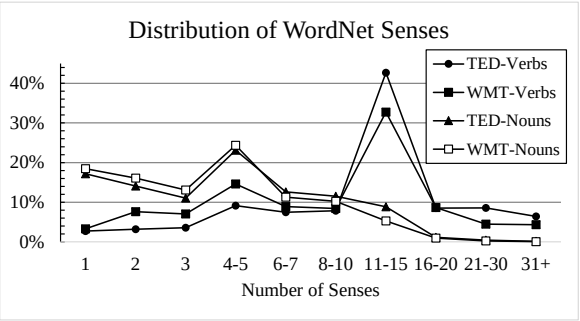


Figure 4: Average lexical translation entropy on English noun and verb stems, computed from the top 95% threshold in the lexical translation table generated by MGIZA.



(a) Average number of senses per verb/noun for the 100 most frequent words in each corpus, as well as the types shared in common (Intersect), and those unique to the respective corpus (Unique-TED and Unique-WMT).



(b) Distribution of WordNet senses for all nouns and verbs in TED and WMT, weighted by observation frequency. Frequencies are bucketed to highlight differences between the corpora.

Figure 3: Polysemy statistics on TED and WMT corpora. Statistics are computed on the 100 most frequent verb/noun stems from each corpus as well as the full list of verbs/nouns found in WordNet.

and target language often results in the mistranslation of pronouns. For example, German has four personal pronoun cases, while English only has two. In cases of high ambiguity, it is up to models that depend on local context, such as n-gram LMs to determine the correct pronoun to use in the translation. If the local features of the sentence cannot resolve the ambiguity, the output pronoun is up to chance. We highlight two additional problems outlined by Hardmeier (2012): the difficulty for anaphora resolution systems to resolve pronouns (e.g. expletive pronouns), and translation divergences, such as when a pronoun is replaced with its referent in the translation.

Using the POS tags assigned by TreeTagger, we identify the English and German pronouns for TED and WMT and report statistics in Table 3. TED contains three times as many pronouns than WMT. While WMT contains few first and second person anaphoric mentions, TED consists of talks in which the speaker often refers to himself and to the audience. In particular, TED and WMT share seven pronominal translations for the English pronoun "you", based on the context of the sentence. At times, "you" may be translated as an indefinite pronoun ("man", "jemand", "eine"), or can be replaced with a different grammatical person ("wir", "sie"). TED contains additional ambiguity which may be attributed to word alignment errors, resulting in high translation entropy (1.53 bits). Likewise the indefinite and ambiguous pronoun "it" occurs twice as often in TED.

Person	Pronouns	TED	WMT	Diff	Rel Diff
1st	10	3.85%	0.48%	3.37%	699.2%
2nd	4	1.68%	0.06%	1.63%	2776.5%
3rd	24	4.06%	2.56%	1.50%	58.6%
Total	38	9.59%	3.10%	6.49%	209.5%

Table 3: Percent of English pronoun tokens in the 2 million word TED and WMT samples. Pronouns are grouped by grammatical person.

Field	TED	WMT	Diff	Rel Diff
Idioms/1K	1.541	2.122	-0.581	-27%
Avg. Length	2.896	2.695	0.201	7.46%
Types	494	556	-62	-11%
Singletons	289	271	18	7%

Table 4: The average rate of idioms per 1,000 words, idiom length, and the number of idiom and singleton types in each corpus sample.

6.4 Idiomatic expressions

Low frequency idiomatic expressions pose challenges for SMT systems. We crawled a list of English idioms generated by an online user community[3]. We manually scanned and pruned a handful of submitted entries that were likely to suggest more false positives than actual idiomatic expressions. In total, we collected 3,720 distinct idiomatic expressions. We perform a greedy idiom search on the surface representation of each corpus, favoring long idioms and ensuring that idioms did not overlap one another. Some statistics are reported in Table 4.

TED and WMT share 237 idioms in common, such as "at the end of the day", "in the face of", and "on the table". These signify expressions that cross genres and are likely to be easily represented with statistical models. Some TED-specific expressions include "beeline for", "bells and whistles", "up the wall", and "warm and fuzzy" – expressions that may be difficult to translate in MT systems trained on news genres. While TED uses fewer idioms overall, nearly 60% of the idiom types appear only once, compared to nearly 50% in WMT.

[3]http://www.usingenglish.com/reference/idioms/

7 Word reordering

One of the most notorious problems in phrase-based statistical machine translation is word reordering (Birch et al., 2009). Expressing the reordering problem as a task of searching through a set of word permutations for a given source sentence **f**, we arrange each source word f_i according to the mean of the target positions $\bar{a_i}$ aligned to it, as suggested by Bisazza and Federico (2013). Unaligned words are assigned the mean of their neighboring words' alignment positions. We then compute a word-after-word distortion length histogram between adjacent source words in their projection to the target language (Brown et al., 1990). To eliminate the effects of sentence length, we randomly sample 100 sentences with replacement for each observable sentence length in each corpus. A histogram is computed for each sentence length, whose results are averaged together.

Figure 5 compares the reordering behaviors of TED and WMT after stratified random sampling. Word permutations are computed from the symmetrized word alignments on English and German stems, using the grow-diag-final-and heuristic in Moses. To visualize the results better, we consider the absolute value of the relative distortion positions. In the figure, Bucket #1 corresponds to discontiguous reordering jumps one position forward (i.e. e_i -- e_{i+1}) or backward (i.e. e_{i+1} e_i), and so on. For example, "we could communicate" is translated once as "wir kommunizieren können" and yields reordering jumps of (+1,-1), which are both binned into Bucket #1. For English-German, monotonic reorderings account for 70.73% and 66.63% for TED and WMT, respectively. This 4% absolute increase in monotonic reorderings is accounted for by the reduction in long distance reorderings of four positions or more.

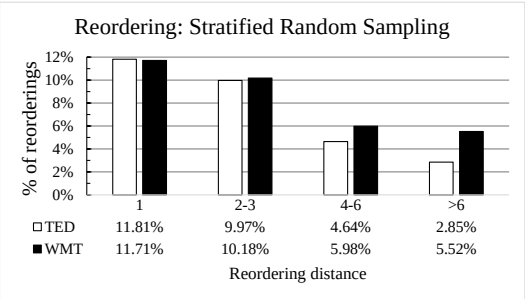


Figure 5: Discontiguous word reordering percentage by reordering distance for English-German. Statistics are computed on reordering buckets of ± 1, $\pm[2,3]$, $\pm[4,6]$, and $\pm[7,\infty)$.

8 Machine Translation performance

Thus far, we have identified several linguistic factors that distinguish the TED translation task from that of WMT News Commentaries. We continue our analysis with a head-to-head comparison of MT performance. Since we cannot directly compare BLEU scores from the two official evaluation tasks, we create a small scale baseline evaluation that fixes the corpora sizes. Using the same two million word samples, we train separate SMT systems on TED and WMT, and tune two held-out samples of 100,000 words. We average the results of three MERT runs to reduce random effects. Each phrase-based SMT system is trained with the default training parameters of Moses (Koehn et al., 2007). We construct separate 4-gram LMs on the German side of the training data with IRSTLM, using a similar configuration as in Section 5.2. To evaluate, we control the effects of sentence length by focusing on sentences containing between 10 and 20 words (after tokenization). For each unique sentence length, we sample 200 sentences with replacement from 300,000 word segments of the TED and WMT corpora. We evaluate using the Translation Edit Rate (TER) metric (Snover et al., 2006). Results are reported in Figure 6 for SMT systems trained with 500K, 1M, and 2M words.

Due to the limited amount of TED data, we cannot measure the effects of additional training data on translation quality, but we attempt to extrapolate the learning curve by looking at smaller training sets. While we cannot explicitly say that TED translation yields higher translation quality than that of WMT, we do observe a growth in the absolute TER difference from 6.4% to 6.8% with 500K words and 2M words, respectively. Likewise, TED has fewer phrase table entries (3.5M vs. 3.7M) and LM entries (1.68M vs. 1.91M 4-grams) than WMT. These results suggest that the characteristics of TED allow better modeling of the translation task with less training data.

9 Discussion

Both TED and WMT News Commentary are good sandboxes for evaluating specific aspects of MT. Our experimental results identify several distinct linguistic phenomena that distinguish each genre's usefulness on specific areas of MT research.

TED talks enjoy performance advantages due to a SMT system's ability to translate their content reasonably well with a surprisingly small amount of training data. While TED has lower lexical ambiguity than WMT in terms of translation en-

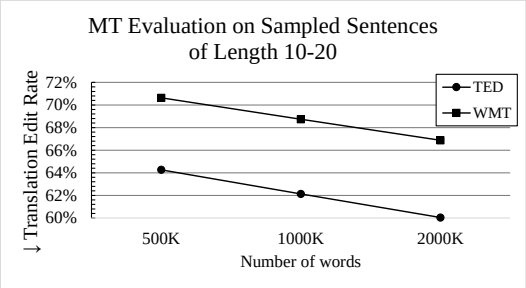


Figure 6: Phrase-based MT results for sampled sentences of length 10-20 in TED and WMT. SMT systems are trained with 500K, 1M, and 2M words.

tropy, it uses significantly more common and thus more ambiguous expressions. Because of this, it is a good candidate for evaluating semantically-informed translation models. The key issue for TED talks is the problem of pronominal anaphora. With over three times as many pronouns than WMT and twice as many third person mentions, the ability for MT systems to handle context is crucial. This makes it is an excellent task for investigating the translation of anaphoric expressions through discourse-aware MT, while at the same time managing the complexity of the system.

As WMT consists of longer sentences with more frequent cases of long distance reordering, it is a better task for measuring differences between hierarchical and linear phrase-based SMT. Additionally, with a lower German-English sentence length ratio, noun and verb compound detection may be a larger issue in WMT. WMT also suffers from higher perplexity scores than TED, suggesting that it can be a good benchmark for evaluating language modeling strategies with large amounts of readily-available in-domain data. Both TED and WMT are good candidates for research on handling idiomatic expressions during translation.

Some linguistic features do not correspond well with the problem of translation difficulty. As shown with our comparison of WordNet polysemy and lexical translation entropy, the challenge of disambiguating between a high number of noun and verb senses lessens during the word alignment process. This could be one of the reasons why previous work on word sense disambiguation in MT has yet to achieve significant improvements in automatic evaluations (Carpuat and Wu, 2007).

It should also be mentioned that while TED appears to be a simpler MT task overall, we have not addressed the larger problem of TED talk translation: the integration with automatic speech recognition. The linguistic features of TED make it a perfect candidate for speech translation, allowing researchers to focus on problems of translating content that may have been corrupted by speech recognition errors.

10 Conclusion

We have shown that the TED spoken language corpus and WMT News Commentary machine translation corpora exhibit differences in several linguistic features that each warrant dedicated research in machine translation. By sampling two million words from TED and WMT, we compared the two corpora on a number of linguistic aspects, including word statistics, such as sentence length and language model perplexity, lexical ambiguity, pronominal anaphora, idiomatic expressions, and word reordering. We observe that while TED consists of shorter sentences with less reordering behavior and stronger predictability through language model perplexity and lexical translation entropy, it has increased occurrences of pronouns that may refer to antecedents in the transcript and a high amount of polysemy through common verbs and nouns. In a small MT experiment, we evaluated a subset of sentence lengths in TED and WMT with MT systems trained on a comparable amount of data and show that TED can be modeled more compactly and accurately.

Finally, we have outlined linguistic features that distinguish the two corpora and propose suggestions to the MT community to focus their attention on TED or WMT, depending on their research goals. While both tasks are interesting for MT research, characteristics of spoken versus written texts provide different challenges to overcome.

Acknowledgments

This work was supported by the EU-BRIDGE project (IST-287658), which is funded by the European Commission under the Seventh Framework Programme for Research and Technological Development. The authors would like to thank Christian Girardi for his assistance with extracting idioms from the web, as well as Arianna Bisazza, Mauro Cettolo, and Carlo Strapparava for other scripts and valuable discussions.

References

D. Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

A. Birch, P. Blunsom, and M. Osborne. 2009. A quantitative analysis of reordering phenomena. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205, Morris-

town, NJ, USA. Association for Computational Linguistics.

A. Bisazza and M. Federico. 2013. Dynamically Shaping the Reordering Search Space of Phrase-Based Statistical Machine Translation. *Transactions of the Association for Computational Linguistics*, 1:327–340.

O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. Rossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

M. Carpuat and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72.

M. Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. In *Proc. of the International Workshop on Spoken Language Translation*, December.

M. Federico, N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech*, pages 1618–1621, Brisbane, Australia.

C. Fellbaum, editor. 1998. *WordNet: an electronical Lexical Database*. MIT Press, Cambridge, MA.

G. Foster, P. Isabelle, and R. Kuhn. 2010. Translating structured documents. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*.

A. C. Graesser, M. Singer, and T. Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychol Rev*, 101(3):371–395, July.

A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202, May.

C. Hardmeier and M. Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the Seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

C. Hardmeier. 2012. Discourse in Statistical Machine Translation. *Discours*, (11), December.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

M. M. Louwerse, P. M. McCarthy, D. S. McNamara, and A. C. Graesser. 2004. Variation in language and cohesion across written and spoken registers. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 843–848.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19:313–330.

J. Mitchell, M. Lapata, V. Demberg, and F. Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206, Stroudsburg, PA, USA.

N. Newbold and L. Gillam. 2010. The linguistics of readability: The next step for word processing. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 65–72, Stroudsburg, PA, USA.

S. T. Piantadosi, H. Tily, and E. Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280 – 291.

A. Schiller, S. Teufel, C. Stöckert, and C. Thielen, 1995. *Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS*. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, August.

L. Specia, N. Hajlaoui, C. Hallett, and W. Aziz. 2011. Predicting Machine Translation Adequacy. In *Machine Translation Summit XIII*, pages 73–80.