Speech to speech machine translation: Biblical chatter from Finnish to English

David Ellis **Mathias Creutz** **Timo Honkela** **Mikko Kurimo**
Brown University
Providence, RI 02912
Helsinki University of Technology
FIN-02015 TKK, Finland

Abstract

Speech-to-speech machine translation is in some ways the peak of natural language processing, in that it deals directly with our original, oral mode of communication (as opposed to derived written language). As such, it presents challenges that are not to be taken lightly. Although existing technology covers each of the steps in the process, from speech recognition to synthesis, deriving a model of translation that is effective in the domain of spoken language is an interesting and challenging task. If we could teach our algorithms to learn as children acquire language, the result would be useful both for language technology and cognitive science.

We propose several potential approaches, an implementation of a multi-path model that translates recognized morphemes alongside words, and a web-interface to test our speech translation tool as trained for Finnish to English. We also discuss current approaches to machine translation and the problems they face in adapting simultaneously to morphologically rich languages and to the spoken modality.

1 Introduction

Effective and fluent machine translation poses many challenges, and often requires a variety of resources. Some are language-specific, some domain-specific, and others manage to be relatively independent (one might even say context-free), and thus generally applicable in a wide variety of circumstances. There are still untapped resources, however, that might benefit machine translation systems. Most statistical approaches do not take into account any similarities in word forms, so words that share a common root, (like "blanche" and "bianca", meaning "white" in French and Italian respectively) are no more likely to be aligned than others (like "vache" and "guardare", meaning "cow" and "to watch" respectively). Such a root is sometimes subject to vowel shift and consonant gradation, and may not be reflected in orthography, since it is often purely phonetic.

This means we are not taking advantage of everything that normally benefits human speakers, hearers and translators. It may be that a more natural approach to translation would first involve understanding of the input, stored in some mental representation (an interlingua), and then generation of an equivalent phrase in the target language, directly from the knowledge sources.

In order to allow for more dramatic differences in grammar like agglutinativity, it seems that the statistical machine translation (SMT) system must be more aware of sub-word units (morphemes) and features (phonetic similarity). This general sort of morphological approach could potentially benefit any language pair, but might be crucial for a system that handles Finnish, Turkish, Hungarian, Estonian or any other highly inflectional language. In the following section we discuss the confounds presented by agglutinative languages, and how awareness of morphemes might improve the situation. This is followed by a brief foray into semantics and natural language generation as a component of

Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 123–130,
Hyderabad, India, January 2008.

SMT. Capturing phonetic features is most applicable to speech-to-speech translation, which will be discussed in the penultimate section. A description of the Bible conversation experiment and some of its results can be found in the final section.

2 Agglutinative Confounds

Traditional n-gram language models and phrase-based translation models do not work terribly well for Finnish because each lexical item can appear in dozens of inflected or declined forms. If an SMT system is presented with "taloosi" (to your house), it will not know if that is another form of a word it saw in training (like "taloissaan", in their houses). Alignment data are thus unnaturally sparse and test sentences often contain several unknown items, which share their stems with trained words. It has been assumed that morphological analysis would be essential for handling agglutinative languages. However, although several effective segmenters and analyzers for specific languages exist, and even unsupervised language-neutral versions such as Morfessor (Creutz and Lagus, 2007), only recently have similar approaches been successfully used in the context of machine translation to improve the BLEU score (Oflazer and El-Kahlout, 2007), and none yet in Finnish.

In our experience, building a translation model through stemmed (truncated) word-alignment outperforms full-form alignment, or any morph-segmented alignment. But once one has generated such a translation model, including phrase tables where stemmed forms (keys in source language) are associated with full forms (values in target language), is there anything to be gained from induction of morphology? Our research in this area has yet to reveal any positive results, but we are still working on it. It is also worth considering the effectiveness of the evaluation metrics. Does BLEU accurately capture the accuracy of a translation, particularly in an agglutinative language? Unfortunately not.

We think the word segmentation in the BLEU metric is biased against progress in morpheme-level translation. Some other metrics have been set forth, but none is widely accepted, in part due to inertia, but also because translation cannot be objectively evaluated, unless both the communicative intent and its effectiveness can be quantified. The same problem occurs for teachers grading essays — what was the student intending to convey, was the phrasing correct, the argument sound, and where does all this diverge from the underlying power of words, written or well said, to transmit information? Translation is an art, and maybe in addition to human evaluation by linguists and native speakers of the language, we should consider the equivalent of an art or literary critic. On the other hand, that might only be worthwhile for poetry, wherein automated translation is perhaps not the best approach.

One might think that the stemmed model described above would lose track of closed-class function items (like prepositions), particularly when they are represented as inflectional morphemes in one language but as separate words in the other. However, it seems that the language model for the target takes care of that quite well in most cases. There are some languages (like Japanese) with underspecified noun phrases, in which efforts to preserve definiteness (i.e., the book, kirjan; a book, kirjaa) seem futile, but given the abundance of monolingual data to train LM's on, these are contextually inferred and corrected at the tail end of the production line. Agglutinative confounds are thus very closely related to other issues found throughout machine translation, and perhaps an integrated solution (including a new evaluation metric) is necessary.

3 Knowledge-Based Approaches

Incorporating statistical natural language generation into a machine translation system involves some modifications to the above. First, the source language is translated or parsed into ontological representations. This is similar to sentence parsing techniques that can be used to induce a context-free grammar for a language (Charniak, 1997), and could in fact be considered one of their more useful applications. The parsing generally depends on a probabilistic model trained on sentences aligned with their syntactic and semantic representations, often in a tree that could be generated by a context-free grammar. The resulting semantic representation can then be used as the source of a target-language generation process.

The algorithm that generates such a representa-

tion from raw input could be trained on a treebank, and an annotated form of the same corpus (where the derivations in the generation space are associated with counts for each decision made) can be used to train the output component to generate language. (Belz, 2005) To incorporate the statistical component, which allows for robust generalization, per (Knight and Hatzivassiloglou, 1995), the NLG on the target side is filtered through a language model (described above). This helps address many of the knowledge gap problems introduced by linguistic differences or in a component of the system - the analyzer or generator.

This approach does have significant advantages, particularly in that it is more focused on semantics (as opposed to statistical cooccurrence), so it may be less likely to distort meaning. On the other hand, it could misinterpret or miscommunicate (or both), just like a human translator. Perhaps the crucial difference is that, while machine learning often has little to do with our understanding of cognitive processes, this sort of machine translation has greater potential for illuminating mysterious areas of the human process. It is not an ersatz brain, nor neural network, but in many ways it has more in common with those technologies (particularly in that they model cognition) than many natural language processing algorithms. That is because, if we can get a semantically-aware machine translation system to work, it may more closely mirror human cognition. Humans certainly do not ignore meaning when they translate, but today's statistical machine translation has no awareness of it at all.

Potential disadvantages of the system include its dependence on more resources. However, this is less of a problem with WordNet(Miller, 1995) and other such semantic webs. It is also worth mentioning again that humans always have an incredible amount of information at their disposal when translating. Not only all of their past experience and word-knowledge, but their interlocutor's demeanor, manner, intonation, facial expressions, gestures, and so on. There are often things that would be obvious in the context of a conversation, but are missing from the transcribed text. For instance, the referent of many pronouns is ambiguous, but usually there is a unique individual or item picked out by the speakers' shared information. This is true for simple sentences like "He hit him," which are normally disambiguated by conversational context, but a purely statistical, pseudo-syntactic interpretation would get little of the meaning a human would glean from that utterance.

4 Spoken Features

Speech-to-speech machine translation is in some ways the peak of natural language processing, in that it deals directly with our (humans') original, oral mode of communication (as opposed to derived written language). As such, it presents challenges that are not to be taken lightly. Much of the pipeline involved is at least relatively straightforward: acoustic modeling and language modeling on the input side can take advantage of the latest advances without extensive adaptation; similarly, speech synthesis on the output can be directly connected with the system (i.e., not work with text output, but a richer representation).

Although such a system might seem quite complicated, it could better take advantage of all the available data. Natural language understanding and generation could even be incorporated to an extent, perhaps to add further confidence measures based on semantic equivalence. Designing it in this way also allows for a variety of methods to be tried with ease, in a modular fashion. It may be that yet another source of information can be found to improve the translation by adding features to the translation model — perhaps leveraging multilingual corpora in other languages, segmenting into morphemes earlier in the process, or even incorporating intonation in some fashion. Weights for all such features could be learned during training, such that no language-specific tuning would be necessary. This framework would certainly not make speech-to-speech translation simple, but its flexibility might make research and improvement in this area more tractable.

Efficiency is crucial in online translation of conversation, so a word alignment model with collapsed Gibbs sampling, rather than EM, at its core is worth experimenting with. We have written up a barebones IBM Model 1 in both C++ and Python, using the standard EM approach and a Gibbs sampling one. The latter allows for optimizations using linear algebra, and although it does not quite match the

perplexity or log-likelihood achieved by EM, it is significantly faster, particularly on longer sentences. Since morpheme segmentation is at least somewhat helpful in speech recognition (Creutz, 2006; Creutz et al., 2007), it should still be considered a potential component in speech-to-speech translation. In terms of incorporating the knowledge-based approach into such a system, we think it may yet be too early, but if existing understanding-and-generation frameworks for machine translation could be adapted to this use, it could be very fruitful, in particular since spoken language generation might be more effective from a knowledge base, since it would know what it was trying to say, instead of relying on statistics alone, hoping the phonemes end up in a meaningful order.

The critical step of SST is, of course, translation. In an integrated system, as described above, the translation model could be trained on a parallel spoken corpus (perhaps tokenized into phonemes, or segmented into morphemes), since there might be advantages to limiting the intermediate steps in the process. The Bible is a massively multilingual publication, and as it happens, its text is available aligned between Finnish and English, and it is possible to find corresponding recordings in both languages. So, this corpus would enable a direct approach to speech-to-speech translation. Alternatively, one could treat the speech recognition and synthesis as distinct from the translation, in which case text corpora corresponding to the style and genre of speech would be necessary. This would be particularly feasible when, for instance, translating UN parliamentary proceedings from a recording, for which translated transcripts are readily available. For a more general and robust solution, we might advocate a combined approach, in the hope that some potential weaknesses of one might be avoided or compensated for by using whatever limited resources are available to add features from the other. Thus, a direct translation from speech to speech could be informed, in a sense, by a derived translation from the recognized text.

5 Biblical Chatter

Here, we present a system for translating Finnish to English speech, in a restricted and ancient domain: the Bible.

5.1 Introduction

Speech to speech translation attacks a variety of problems at once, from speech recognition to synthesis, and can similarly be used for several purposes. If a system is efficient enough to be used without introducing significant delay, it can translate conversational speech online, acting as an interpreter in place of (or in cooperation with) a human professional. On the other hand, a slow speech translation system is still useful because it can make news broadcasts (radio or television) accessible to wider audiences through offline multilingual dubbing, allowing international viewers to enjoy a delayed broadcast.

5.2 System Description

The domain selected for our experiments was heavily influenced by the available data. We needed a bilingual (Finnish and English) and bimodal (text and speech) corpus, and unfortunately none is readily available, but we put one together using the Bible. Both Old and New Testaments were used, with one book from each left out for testing purposes. We used multiple editions of the Bible to train the translation model: the American Standard Version (first published in 1901, updated 1997), and Finnish translations (from 1992 and 1933,38). The spoken recordings used were the World English Bible (1997) and Finnish Bible (Raamattu) readings (recorded at TKK 2004).

Our approach was to use existing components, and try weaving them together in an optimal way. First, there is the open vocabulary automatic speech recognition (ASR) task, where the goal is to detect phonemes in an acoustic signal and map them to words. Here, we use an "unlimited vocabulary" continuous speech recognizer (Hirsimäki et al., 2006), trained on a multi-speaker Finnish acoustic model with a varigram (Siivola et al., 2007) language model that includes Bible n-grams. Then, for translation, Moses (Koehn et al., 2007) is trained on words and morphemes (derived from Morfessor Baseline (Creutz and Lagus, 2005)). For speech synthesis, we used Festival (Taylor, 1999), including the built-in English voice and a Finnish voice developed at Helsinki University.

5.3 Results

The following is an example fragment, taken from the test corpus.

Niin Daavid meni	David slept with his
lepoon isiensä luo,	fathers, and was
ja hänethaudattiin	buried in the
Daavidin kaupunkiin.	city of David. The days
Neljäkymmentä vuotta	that David reigned
hän oli ollut	over Israel were
Israelin kuninkaana.	forty years; seven
Hebronissa hän	years reigned he
hallitsi seitsemän	in Hebron, and
vuotta, Jerusalemissa	thirty-three years
kolmenkymmenenkolmen	reigned he
vuoden ajan.	in Jerusalem.
Salomo nousi	Solomon sat on
isänsä Daavidin	the throne of David
valtaistuimelle,ja	his father; and
hänen kuninkuutensa	his kingdom was
vahvistui lujaksi.	established greatly.

A translation of the reference text skips recognition, and runs the system from translation to synthesis. The following shows how the sample text was translated by our system (BLEU = 0.735):

Niin Daavid meni	so david slept with his
lepoon isiensä luo,	fathers and was
ja hänet haudattiin	buried in the
Daavidin kaupunkiin.	city of david
Neljäkymmentä vuotta	forty years he
hän oli ollut	was king over
Israelin kuninkaana.	israel and in
Hebronissa hän	hebron he reigned
hallitsi seitsemän	seven years
vuotta, Jerusalemissa	in jerusalem
kolmenkymmenenkolmen	thirty and three
vuoden ajan.	years solomon
Salomo nousi	went up to
isänsä Daavidin	the throne of
valtaistuimelle, ja	david his father
hänen kuninkuutensa	and his kingdom
vahvistui lujaksi.	was strong for luja

The following recognized translation (BLEU = 0.541) represents a complete run of the system. The recognition (on the left) had a LER of 12.9% and a WER of 56.8%.

niintaa meni	niintaa went
lepoon isiensälla	isiensälla rest and was
ja hänet haudattiin	buried in the
daavidin kaupunkiin	city of david the king
neljäkymmentä	of israel was
vuotta hän oli ollut	forty years he was
israelin kuninkaan	in hebron he
hebronissa hän	reigned seven years
hallitsi seitsemän	in jerusalem
vuotta jerusalemissa	kymmenenkolmen
kolmen kymmenenkolmen	three years
vuoden ajan	after the new
salomon uusi	on the throne of david
isänsä daavidin	and solomon
valtaistuimelle ja	his father
hänenkuninkuutensa	hänenkuninkuutensa
valmistulujaksi	valmistulujaksi

Here we have an alternative path through the system, which uses Morfessor on the recognized text, and then translates using a model trained on the morpheme-segmented corpus. This results in a reduced score (BLEU = 0.456), but fewer unknown words.

n iin taa# meni#	iin behind went to
lepo on# isi ensä lla#	the sabbath that
ja# hän et# hauda ttiin#	is with ensä and he
daavid in# kaupunki in#	shall not at the grave of abner
neljäkymmentä# vuotta#	was forty years of the
hän# oli# ollut#	city of david and
israeli n kuninkaan#	he was israeli to
hebron issa# hän#	the king of hebron
hallitsi# seitsemän#	and he reigned
vuotta# jerusalem	seven years in
issa# kolmen#	jerusalem three tenth
kymmenen kolmen#	three years of
vuoden# ajan#	the new solomon his
salomo n# uusi# isä	istuim to david
nsä# daavid in#	my father of the
valta istuim elle#	kingdoms of
ja# hän en kun ink	ink and he
uutensa# valmistu luja ksi#	uutensa valmistu to luja

The morphemes might have been more effective in translation if they had been derived through rule-based morphological analysis. Or, they could still be statistical, but optimized for the translation phase by minimizing perplexity during word alignment.

A significant barrier to thorough and concrete evaluation of our system involves segmentation of the speech stream into sentences (or verses) to match the text. In the above examples, we manually clipped the audio files. Evaluating performance on the entire test set reduced the BLEU score if the data were streamed through each component unsegmented. When the recognizer was set to insert a period for detected pauses of a certain length, or at sentence boundaries identified by its language model,

input to the translation phase became considerably more problematic. In particular, the lattice input ought to be split into sentences, but there would usually be a period in every time slice (but with low probability).

5.4 Discussion

There were significant difficulties in the process, particularly in the English to Finnish direction. Whereas Finnish speech recognition is relatively straightforward, since its orthography is consistent, English speech recognition is more dependent on a pronunciation dictionary. Although many such dictionaries are available, and the pronunciation of novel words can be estimated, neither of these resources is terribly effective within the Bible domain, where there are many archaic words and names. In the second step, translation into Finnish is demonstrably difficult from any source language, and results in consistently lower BLEU scores (Virpioja et al., 2007). And although using morphemes can reduce the frequency of unknown words, it also reduces the BLEU score.

It might improve translation quality if we use the recognizer lattice as translator input, since acoustically improbable segments may lead to the most fluent translation. Having access to many possibilities might help the translation model, but then again, second-guessing the recognizer might not be helpful. There were some difficulties with the Moses integration, in part because the word-sausage format varies from SRILM's. Also, the recognizer output indicates word boundaries as <w>, not string-final hash-marks (#). This is problematic since the former are separate symbols, occupying a node in the lattice, whereas the latter are appended to another symbol (e.g., "<w> morph eme </w>", 4 nodes, versus "morph eme#", 2 nodes). Using the lattice, final output from Moses tends to be more fluent, but less on-topic, and often truncated. Although we have no improvements thus far, it is likely that with further parameter tuning, we could achieve better results. On the other hand, we seek a general, robust, domain-independent solution, so focusing on Bible translation may not be worthwhile.

Our speech-to-speech translation system is accessible through a web interface.

`http://www.cis.hut.fi/projects/speech/translate/`

It accepts a sound file, with recorded Finnish bible-style chatter, an optional reference text and translation, and within a half hour (usually much less) sends a detailed report, including a sound file with the synthesized English.

Ideas for future research include online speech-to-speech translation, which must be efficient, lightweight and robust. A potential barrier to this and other applications is the lack of spoken language training texts. It might be possible to adaptively train to new speakers and contexts, perhaps taking advantage of an efficient alternative to EM in word alignment (see discussion of Gibbs sampling). As mentioned elsewhere, it might be worth using prosodic features captured during recognition as factors in translation. Adapting existing resources to new language pairs is particularly essential in an area where so much is necessary, and so little available.

6 Conclusion

We cannot yet say for sure whether linguistic or statistically optimized morphemes derived from text corpora could be useful somehow in machine translation, but it has been demonstrated helpful in speech recognition. Awareness of sub-word units could benefit a speech-to-speech translation system, and it may in fact help to maintain information from the speech recognizer about morpheme segmentation throughout the translation process, even in speech generation. Incorporating natural language understanding may also be fruitful, but for compact, efficient systems (like a handheld translator) might not have access to the necessary resources or computational power to support that. On the other hand, it is our duty as researchers to stay ahead of the technology and push its limits.

We are by no means the first to imagine this, but perhaps we will soon be speaking into wrist watches that understand our query, seemingly instantly shift through more information than Google has currently indexed, and reply in fluent English, Finnish, or Punjabi with as much detail as could be hoped for after hours of painstaking research with current technology. In this case (and computational linguists must always be optimistic), knowledge-based natural language processing certainly has a crucial place.

Morphemes and agglutinative languages do pose unique problems for computational linguists, but many of the general techniques developed for languages like Arabic and Chinese, which are equally far from English in grammar (and even orthography), might surmount those problems without any manual adaptation. Discriminative training of features used in the translation model allows for such solutions to be molded automatically to whatever language pair (and set of corpora) they are being used for. There is, as always, much more to be done in this area, and we hope our research into efficient, online Bible-conversational translation — a modern Babelfish in an ancient genre — is fruitful, and helps to shed light on lemmatization.

Acknowledgments

Many thanks to Teemu Hirsimäki, Antti Puurula, Sami-Virpioja and Jaakko J. Väyrynen for their help with components of the system and for their thoughts and comments at various stages of the project.

References

Anja Belz. 2005. Statistical generation: Three methods compared and evaluated. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG05)*, pages 15–23.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *AAAI/IAAI*, pages 598–603.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1), January.

Mathias Creutz, Teemu HirsimŁki, Mikko Kurimo, Antti Puurula, Janne Pylkknen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraclar, and Andreas Stolcke. 2007. Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. In *Proceedings of Human Language Technologies / The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007)*, Rochester, NY, USA.

Mathias Creutz. 2006. Morfessor in the morpho challenge. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, Venice, Italy.

T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541.

Kevin Knight and Vasileios Hatzivassiloglou. 1995. Two-level, many-paths generation. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 252–260, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar, Alexandra Constantin, and Evan Herb. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180.

George A. Miller. 1995. Wordnet: a lexical database for English. *Commun. ACM*, 38(11):39–41.

Kemal Oflazer and Ilknur Durgar El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 25–32.

Vesa Siivola, Teemu Hirsimäki, and Sami Virpioja. 2007. On growing and pruning Kneser-Ney smoothed n-gram models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5):1617–1624.

Paul Taylor. 1999. The Festival Speech Architecture. Web page.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology–aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of the Machine Translation Summit XI*, Copenhagen, Denmark. To appear.