Corpus Support for Machine Translation at LDC

Xiaoyi Ma, Christopher Cieri

Linguistic Data Consortium
3600 Market St. Suite 810
Philadelphia, PA 19104
{xma, ccieri}@ldc.upenn.edu

Abstract

This paper describes LDC's efforts in collecting, creating and processing different types of linguistic data, including lexicons, parallel text, multiple translation corpora, and human assessment of translation quality, to support the research and development in Machine Translation. Through a combination of different procedures and core technologies, the LDC was able to create very large, high quality, and cost-efficient corpora, which have contributed significantly to recent advances in Machine Translation. Multiple translation corpora and human assessment together facilitate, validate and improve automatic evaluation metrics, which are vital to the development of MT systems. The Bilingual Internet Text Search (BITS) and Champollion sentence aligner enable the finding and processing of large quantities of parallel text. All specifications and tools used by LDC and described in the paper are or will be available to the general public.

1. Introduction

Despite recent advances, Machine Translation (MT) remains one of the most difficult tasks in human language technology. The difficulty results not only from the complexity of human language but also from the lack of resources to try out approaches other than traditional rule-based MT, which requires thousands of human labor years to craft tens of thousands of rules and seems to have reached a plateau in its performance.

At the beginning of the DARPA-sponsored TIDES program, which focused on the translation of Arabic and Chinese into English, it was clear that large amount of linguistic data would be required to support approaches such as Statistical Machine Translation (SMT) (Brown et al. 1990; Brown et al. 1993; Vogel and Tribble 2002; Yamada and Knight 2001; Papineni et al. 1998; Och and Ney 2004) to make the program a successful one. These data – lexicons, large parallel text corpora, multiple translation corpora, human assessments, treebanks and monolingual text – had to be collected or created efficiently and at a very low cost, an very ambitious and challenging goal given that, at the time, we had very little data in our possession and our knowledge of how much such data existed and how to create new data was limited.

The approach we took was to maximize the use of existing data and minimize the use of human labor. We created tools for harvesting existing data from the Internet, and we tried to automate as many steps in the collection and processing pipeline as possible. Some of these automated steps not only reduced the time needed to collect the data, but also outperformed humans.

In this following paper, we describe the approaches we took to create lexicons, parallel text corpora, multiple translation corpora and human assessment data. The development of other types of data is covered by other LDC publications, such as (Mammouri and Bies 2004).

2. Lexicons

Our main focus in lexicons are translation lexicons, morphological analyzers and bilingual named entity lists.

2.1. Translation Lexicons

A translation lexicon contains entries of source words and their possible translations in the target language. It differs from a traditional bilingual dictionary in that lacks full form definitions, examples, usage notes and the like.

The Internet provides us with a large number of online bilingual dictionaries that can provide initial input to our lexicon. Some of them may be downloaded with the click of a mouse, but most of them require ad hoc processing to retrieve the entries one by one. In most cases we needed a source word list in English, Chinese or Arabic.

After the online bilingual dictionaries were downloaded, steps were taken to 1) normalize the formats, 2) combine different dictionaries, 3) remove duplicates, and 4) remove bogus entries by first filtering out obvious bad entries using automated methods and then manually examining the remaining entries.

Besides Arabic and Chinese translation lexicons, we also produced translation lexicons in many other languages, including Hindi, Thai, Punjabi.

2.2. Bilingual Named Entity List

Bilingual named entity lists are rare. We were able to obtain a large named entity database created by Xinhua News Agency.

Limited QC was performed on the entire set. The English->Chinese version of each pair was created by reversing the Chines->English, both sorted by the Unix built-in sort function.

The contents are as follows:

Lists	Direction	#entries
Place Names	Chinese to English	276,382
Place Names	English to Chinese	298,993
Organization Names	Chinese to English	30,800
Organization Names	English to Chinese	37,145
Corporate Names	Chinese to English	54,747
Corporate Names	English to Chinese	58,468
Press Organization Names	Chinese to English	29,757
Press Organization Names	English to Chinese	32,922
Intl. Organization Names	Chinese to English	7,040
Intl. Organization Names	English to Chinese	7,040

We plan to compile similar Arabic – English bilingual named entity lists in the near future.

2.3. Buckwalter Arabic Morphological Analyzer

The Buckwalter Arabic Morphological Analyzer is a morphological parser developed by Tim Buckwalter.

Morphotactics and morphophonemic rules were built directly into the three lexicon files for Arabic prefixes, suffixes and stems. The prefix lexicon contains all valid concatenations of prefixes. Similarly, the suffix lexicon contains all valid concatenations of suffixes. The morphophonemic rules were treated simply as orthographic variations and addressed by means of additional dictionary entries. Although an Arabic parser with "unvocalized" lexicon entries can be built and would be fully functional, short vowels and diacritics are included in the lexicons because without them. the lexicon is extremely difficult to maintain and the analysis output is also difficult to interpret.

The analyzer assumes each Arabic word contains a prefix, a stem and a suffix, in which the prefix and suffix can be empty. For each segmentation hypothesis, the analyzer checks against the corresponding lexicon to see if the word element exists. If all three word elements (prefix, stem, suffix) are found in their respective lexicons, their respective morphological categories are then used to determine whether they are compatible. A morphological analysis is valid only if all three word elements exist in their respective lexicons and the elements are compatible.

The analyzer currently contains 548 prefixes, 906 suffixes, and 78,839 stems representing 40,219 lemmas. It uses the Buckwalter Transliteration scheme internally, and it can read input in both Buckwalter Transliteration and Window CP-1256 encoding.

A sample analysis is as follows:

INPUT STRING: والغاز
LOOK-UP WORD: wAlgAz
SOLUTION 1: (wa>alogAz) wa/CONJ+>alogAz/NOUN
(GLOSS): and + mysteries/enigmas/riddles +
SOLUTION 2: (waAlgAz) wa/CONJ+Al/DET+gAz/NOUN
(GLOSS): and + the + gas +

For a more detailed description of the Buckwalter Arabic Morphological Analyzer, please refer to (Buckwalter 2006).

3. Parallel Text

Parallel text is the foundation of a number of promising MT approaches, including Statistical MT and Example Based MT. Rule Based MT also finds parallel text useful for automatic extraction of transfer rules.

We utilize a combination of approaches to collect/create parallel text:

Collecting from known institutions, including the United Nations, the European Union, and the government of Hong Kong Special Administration Region. This is the simplest and most rewarding means of collecting parallel text, because these institutions usually have a large and growing archive of parallel text. Our experiences show that obtaining permission for redistributing the collected data is manageable;

Collecting from the Internet. Though the Internet contains a tremendous amount of data and is a gold mine of parallel text, finding parallel text is not a trivial task because of the sheer size of the Internet and the language expertise required for identifying translated text. We developed a tool called the Bilingual Internet Text Search (BITS) (Ma and Liberman 1999), which scans thousands of websites per day to find parallel text of a specified language pair. BITS requires a translation lexicon, and a light stemmer for morphologically complex languages. A more detailed description of BITS follows.

Outsourcing translation is an approach we can employ at increased cost when insufficient parallel text is recovered through the two approaches discussed above. Although creating new translated text is very expensive and time consuming, it does have a few advantages over collected parallel text: 1) texts from any genre of our choosing can be translated; 2) the translation quality can be controlled; 3) perfect sentence alignment can be achieved by segmenting the source text prior to being translated.

Parallel text provides the maximum utility when it is sentence aligned. Although sentence alignment has been extensively studied and is considered a solved problem by many researchers, it remains a challenge in practice, especially when the parallel text was noisy, which is often the case when we are dealing with a large body of parallel text of different formats and encodings. Our research in sentence alignment produced Champollion (Ma 2006), a lexicon-based sentence aligner, which is very robust on noisy data.

3.1. BITS

BITS is a tool for finding parallel text over the Internet without human intervention. The input of BITS is a list of web sites which possibly contain parallel text of the pertinent language pair. The output is parallel text aligned at document level.

The major components of BITS are:

1) **A Language Identifier** to identify the language in which a text is written. The language identifier is a trainable ngram-based language identifier, which requires about 50K words of monolingual text for each language it identifies. The language identifier can also identify the different encodings of the same language, for instance Windows CP1256 and UTF8 for Arabic text.
2) **A Webpage Retriever** to retrieve web pages recursively from remote web servers. There are several such tools available, including GNU wget and curl. BITS uses wget with a wrapper.
3) **A Document Aligner** to find translated files among files of the specified language pair. Researchers have experimented with URL-based alignment (matching files based on the similarity of their URLs) and markup-based alignment (matching files based on the HTML tags in files), but neither is reliable. The BITS document aligner uses a content-based matching algorithm supported by a translation lexicon. It counts the number of translated words between corresponding regions of the two texts. If the percentage of translated words exceeds a certain threshold, the two texts would be considered a

match. We usually set the threshold high enough to achieve high precision. As a result, the recall would be lower, but we save the trouble of manually checking the matches. For language pairs where translated text is rare, we set the threshold low so that more matches can be identified. The false positives can be eliminated afterwards by eyeballing the harvested text.

BITS takes the following steps to process a website:

1) **Website Language Identification** to identify the languages of a given web site. This is done by first retrieving the top 4 to 5 levels of web pages of the website and then identifying the language of each retrieved page.
2) **Web Page Retrieval** to download the entire website if the languages of the website match the target language pair.
3) **Webpage Cleanup** to remove menu bars, tables, etc. which have negative effects on language identification, and to strip off html tags.
4) **Page Language Identification** to identify the language of each web page.
5) **Document Alignment** to find matches among the pages of the two specified languages.

BITS requires a translation lexicon of 10K entries. We were able to find translation lexicons on the Internet for all the language pairs that we were interested in.

BITS also requires a list of websites as input. There are two ways to create the list: 1) querying Domain Name Servers (DNS), for example, asking a DNS server to list all sub domains of the .in domain; 2) searching for common but distinctive words of a relevant language on commercial search engines, such as Google. Querying DNS provides a more comprehensive list of a domain, but not every DNS server allows unrestricted querying, in which case searching for common words is the only option.

BITS has been very successful locating parallel text, for example, it found over 1M words of Thai – English parallel text over a period of one week. It has failed to find large amount of Urdu – English parallel text, however, probably because there is not much available on the Internet.

3.2. Champollion

Real world parallel text can be very noisy. This is especially true when it comes to processing large archives of parallel text, or parallel text harvested from the Internet. The noise can come from several sources: 1) loose translation – translators deleted or inserted sentences when translating; 2) deletion and insertion resulted from preprocessing – when processing very large archives of data in many different formats, for example, 10 years of UN data, corners have to be cut when preparing the data for sentence alignment, which will certainly lead to the deletion of some text of one language but not the other. As a comparison, the deletions and insertions account for only 1.3% in the UBS corpus (Gale and Church 1991), but more than 6.3% in a corpus we sampled from three large parallel text archives (Ma 2006).

All existing sentence aligners perform well on clean data – (Gale and Church 1991) reports 94.2% accuracy on the English – French UBS data by their pure length-based algorithm. However, their performance on noisy data is far from satisfactory. Starting with Chinese – English, we made significant efforts towards a robust sentence alignment algorithm, which resulted in Champollion, a lexicon-based sentence aligner.

Champollion differs from other lexicon based sentence aligners in two ways. First, it assumes a noisy input, i.e. a significant percentage of non-one-to-one mappings, including deletions and insertions. Champollion considers a sentence pair a possible match only when there are lexical matches, while other approaches may consider it a match based on other information, such as length, which are often unreliable when dealing with noisy data.

Second, unlike other approaches, Champollion assigns greater weight to translated words that are less frequent in the context. For example, assuming we have the following sentence pair in a report on recent waves of violence in Iraq:

a. Marketplace bombing kills 23 in Iraq
b. 伊拉克 集市 爆炸 造成 23 人 死亡

The translation pair (23, 23) is much stronger evidence than (Iraq, 伊拉克) that the two sentences are a match, simply because “Iraq” and “伊拉克” appear much more often than “23”.

Champollion then uses dynamic programming to find the optimal alignment. It allows a 1-1, 1-0, 0-1, 1-2, 2-1, 2-2, 1-3, 3-1, 1-4 and 4-1 alignment. There is a penalty associated with alignments other than 1-1 alignment. The penalty is determined empirically. Sentences with a mismatching length are also penalized.

Experiments show that Champollion precision and recall are slightly better than other aligners on clean data, but much better than the others on noisy data.

Champollion has been used extensively at the LDC on aligning Chinese – English, Arabic – English, and Hindi – English parallel text.

Please refer to (Ma 2006) for a complete description of Champollion.

3.3. Outsourcing Translation

The specification for outsourcing translation evolved over the years as existing procedures were improved and new issues were addressed.

We employ multiple translation teams for each language. Each team is required to have at least one translator native in the source language and one native in the target language. A translation company can have more than one team, working either collaboratively or independently as the projects require. Before a translation team is hired, it is required to provide a good translation of an article of about 250 words to qualify.

The translation guidelines provide detailed instructions and examples with regard to translating proper names, speech disfluencies, factual errors, typos etc. Please refer to (LDC 2006) for the full translation guidelines.

3.3.1. Source Text Format

The original text or speech transcripts the LDC creates or acquires are in many different formats, which, besides text and transcripts, also include metadata such as speaker IDs, section boundaries, turn boundaries and timestamps, etc. The LDC reformats the source text before sending it to the translators to 1) make the source files easy to read;

2) to avoid translator's tampering with metadata; and 3) to aid automatic processing when the translation is returned to LDC.

The source text is sentence segmented before it is translated. The sentence boundaries are kept intact during translation so that the translation and the source text are perfectly sentence aligned. Each source file is formatted as such, taking Arabic speech transcripts as an example:

```
<ar=1> [speaker1] {Arabic sentence 1}
<en=1>
<ar=2> [speaker1] {Arabic sentence 2}
<en=2>
<ar=3> [speaker2] {Arabic sentence 3}
<en=3>
```

A source file contains multiple Arabic lines, each followed by an English line as the placeholder for the English translation of the Arabic sentence. Each Arabic source line contains a sentence ID, a speaker ID and the source text. The speaker IDs are for speech transcripts only; they help translators understand the conversation.

Translators are instructed not to change any part of the source file except typing in the English translation.

3.3.2. Quality Control

LDC's fluent bilingual staff review every delivery. The translation teams are not paid until the translation passes quality control.

For each delivery, we randomly select a subset of the documents, and choose 5 consecutive segments from any part of the files, until the total number of words adds up to about 1,200. The selected sample will then be graded using the system described below.

Error	Deduction
Syntactic	4 points
Lexical	2 points
Poor English usage	1 point
Significant spelling or punctuation error	½ point (to a maximum of 5 points)

For each error found, the corresponding number of points will be deducted. If more than 40 points are deducted from the 1200-word sample, the translation will be considered unacceptable and the whole delivery will be sent back to the translation team for improvement.

Several commercial translation agencies have failed the QC in the past a few years. They either found ways to improve their translation quality, or they were dropped from our vendor list.

In general, we have seen a significant improvement in translation quality since the QC procedure was put into place. The work load for quality control was significant at the beginning, dying down gradually as teams made adjustments to meet our standard of translation quality.

4. Multiple Translation Corpora

Human evaluation of machine translation quality is expensive, time consuming and involves human labor which cannot be reused. Relying only on human assessment would stunt growth in MT because researchers and developers need to monitor the effect of the daily changes they make to their systems so as to weed out the bad ideas from the good ones.

Automatic evaluation metrics have been extensively studied in recent years. Most metrics that are being used are based on IBM-invented BLEU metric (Papineni et al. 2002), which scores MT output by calculating its ngram matches against multiple human translations – that is, translations of the same text by multiple professional translators who work independently. BLEU relies on multiple translations to model the variation among human translators.

We create multiple translation corpora to support the development and evaluation of MT systems. These corpora are being used to develop better automatic evaluation metrics as well.

We use the best teams to produce multiple translation corpora. The teams are instructed to work independently. The names or pseudonyms of the translators, their qualifications, the translation aid they use, if any, and the procedure of their translations are carefully documented.

We use the same quality control procedure described in section 3.3.2 to ensure the quality of the multiple translation corpora.

5. Human Assessment

We conduct human assessment to evaluate the performance of MT systems in comparison with human translation teams and commercially available translation systems, and to see how well automatic evaluation metrics track human judgments.

The human assessment specifications were based on the 1993 DARPA MT human assessment specifications in which translations are evaluated on the basis of adequacy and fluency. Adequacy refers to the degree to which the translation communicates information present in the original or in a best of breed translation that serves as a proxy to the original. Fluency refers to the degree to which the translation is well-formed according to the grammar of the target language.

A team of human judges provide multiple (a minimum of two) assessments of adequacy and fluency for each sampled segment of each translation of each story. For "adequacy" assessments, judges compare each segment to a gold standard. A bilingual linguist chooses the best of the human translations to serve as the gold-standard. Fluency is assessed with respect to the grammar of Standard Written English and requires no comparison. Judges view each translated sentence only once, giving fluency and adequacy assessments in a single pass. Assessment is timed and judges are strongly encouraged to work as quickly as comfortably possible.

For each translation of each segment of each selected story, judges make the fluency judgment before the adequacy judgment.

5.1. Fluency Assessment

A fluent segment is one that is grammatically well formed; contains correct spellings; adheres to the common use of terms, titles and names; is intuitively acceptable; and can be sensibly interpreted by a native speaker of English. A fluency judgment is one of the following:

	How do you judge the fluency of this translation? It is:
5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

5.2. Adequacy Assessment

Having made the fluency judgment for a translation of a segment, the judge is presented with the "gold-standard" translation. Comparing the target translation against the gold-standard, judges determine whether the translation is adequate. Adequacy refers to the degree to which information present in the original is also communicated in the translation. Thus, for adequacy judgments, the gold-standard will serve as a proxy for the original source-language text. An adequacy judgment is one of the following:

	How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?
5	All
4	Most
3	Much
2	Little
1	None

Where English translations retain source words or characters from the original news stories, judges are instructed to give a score between "1: None" and "4: Most" depending upon the degree to which the un-translated words or characters, among the other factors, affect the adequacy of the translation.

5.3. Assessment System

The "assessment system" is defined here as the collection of utilities, computer programs, and graphical user interfaces that prepare the output of the human translation teams for assessment, assign translation to individual human judges, display segments of the translations, collect human judgments on them, and output the human judgments in the output format specific above.

The assessment system distributes translations of the original news stories across judges such that each judge reviews a roughly equal number of translations and such that two independent judges assess each translation of each story. The assessment system presents segments within a story in their naturally occurring order but otherwise provides all translation of all stories in random order. The assessment system ensures that stories and translations of stories are distributed randomly across judges. Specifically, except as may occur in a random sampling, the assessment system does not assign any one judge a disproportionate percentage of either translations of one original story or of translation by a single translator.

The assessment system's GUI presents all segments of a selected translation in the order in which the segments appeared in the original news story. For each selection, the GUI first presents the segment alone and acquires a fluency judgment. The interface then displays the corresponding gold-standard segment and acquires an adequacy judgment before progressing to the next segment. The GUI does not display the gold-standard segment while the judge is making the fluency assessment.

Please refer to (LDC 2005) for a complete human assessment specification.

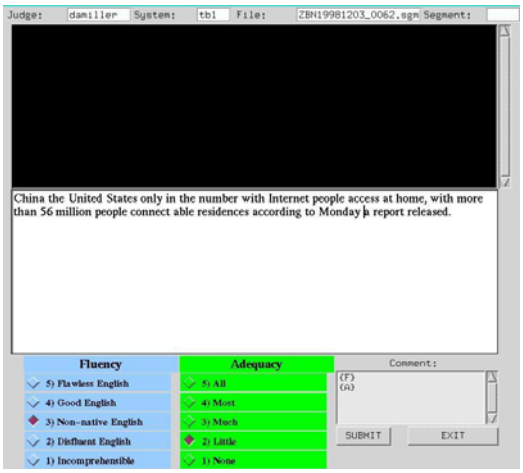

Figure 1: Interface for fluency judgment

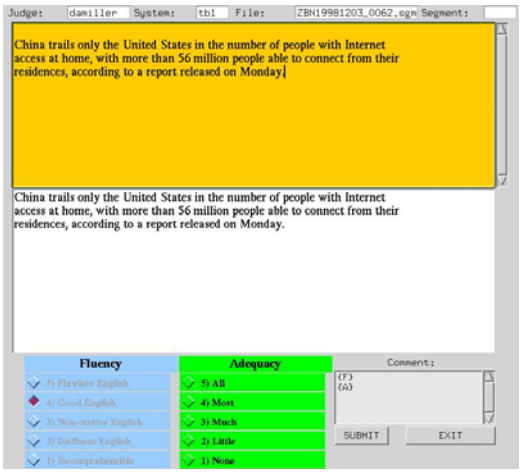

Figure 2: Interface for adequacy judgment

6. Achievements

During TIDES, we created 3 lexicons, 11 parallel text corpora, and 10 multiple translation corpora. Beginning with nothing, we created 210 million words of Chinese – English parallel text and 110 million words of Arabic – English parallel text. Most of these are already available to general public, the remainder is in our publication pipeline.

The tools we developed are being used extensively by LDC projects, such as GALE and Less Commonly Taught Languages, and by other institutions as well.

The following is a list of publish corpora at LDC:

Chinese <-> English Name Entity Lists (v1.0)
Buckwalter Arabic Morphological Analyzer Version 2.0
Hong Kong Parallel Text
Chinese English News Magazine Parallel Text
Chinese News Translation Text Part 1
Arabic News Translation Text Part 1
Arabic English Parallel News Part 1
Multiple-Translation Chinese Corpus
Multiple-Translation Chinese (MTC) Part 2
Multiple-Translation Chinese (MTC) Part 3
Multiple Translation Chinese (MTC) Part 4
Multiple-Translation Arabic (MTA) Part 1
Multiple-Translation Arabic (MTA) Part 2
Chinese Treebank 5.0
Arabic Treebank: Part 1 v 3.0
Arabic Treebank: Part 2 v 2.0
Arabic Treebank: Part 3

The following is a sample of what will be released in the near future:

UN Chinese English Parallel Text
FBIS Multilanguage Texts
UN Arabic English Parallel Text
Arabic Treebank English Translation
Corporate News Parallel Text
Multiple Translation Chinese (MTC) Part 5
Multiple Translation Chinese (MTC) Part 6
Multiple-Translation Arabic (MTA) Part 3
Multiple-Translation Arabic (MTA) Part 4

7. Conclusion

This paper describes a subset of our efforts in supporting the research and development of automatic machine translation. Backed by our experience in data scouting, annotation, and core technology research and management, we are able to provide the MT research community with lexicons, very large parallel text corpora, multiple translation corpora, human assessment of translation quality, treebanks, monolingual text, etc.

These data address many aspects of MT research – training, evaluation, human assessment and automatic evaluation. By utilizing existing resources and automating our collection and processing pipeline, we are able to create these data quickly, cost efficiently, and with high quality.

The machine translation corpora are also being used for researches in information retrieval and language teaching.

8. References

Buckwalter, T. (2006). The Buckwalter Arabic Morphological Analyzer. In *Arabic Computational Linguistics: Current Implementations, CSLI Publications*. [forthcoming]

Brown, P., Cocke, J., Della Pietra, V., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A statistical approach to Machine Translation. In *Computational Linguistics* 16, 2, pp. 79-85.

Brown, P., Della Pietra, V., Della Pietra, S., and Mercer, R. (1993). The mathematics of statistical Machine Translation: Parameter estimation. *Computational Linguistics* 19, 2, pp. 263-311.

Gale, W. A., Church, K. W. (1991). A program for aligning sentences in bilingual corpora', in *ACL '91*, Berkeley CA, pp. 177-184.

LDC (2006). LDC Human Translation Guidelines, *http://projects.ldc.upenn.edu/gale/Translation*.

LDC (2005). Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. *http://projects.ldc.upenn.edu/TIDES/tidesmt.html*.

Ma, X. (2006). Champollion: A Robust Parallel Text Sentence Aligner. In *Proceedings of LREC-2006*, Genoa, Italy.

Ma, X., Liberman, M. (1999). BITS: A Method for Bilingual Text Search over the Web. In *Proceedings of the Machine Translation Summit VII*, Singapore.

Maamouri, M., Bies, A. (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004*, Geneva.

Och, F., Ney, H. (2004). The alignment template approach to statistical machine translation. In *Computational Linguistics*, pp. 417-449.

Papineni, K., Roukos, S., and Ward, T. (1998). Maximum likelihood and discriminative training of direct translation models. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-98)*, pp. 189-192.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, PA, pp. 311-318.

Vogel, S. and Tribble, A. (2002). Improving statistical machine translation for a speech-tospeech translation task. In *Proceedings of the 7th International Conferenc on Spoken Language Processing (ICSLP-02)*.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Anniversary Meeting of the Association for Computational Linguistics (ACL-01)*.