Creating a High-Quality Machine Translation System for a Low-Resource Language: Yiddish

Dmitriy Genzel and **Klaus Macherey** and **Jakob Uszkoreit**
Google, Inc.
1600 Amphitheatre Pkwy
Mountain View, CA 94043, USA
{dmitriy,kmach,uszkoreit}@google.com

Abstract

We introduce the first machine translation system for Yiddish-English and English-Yiddish. We discuss challenges presented by this language and their solutions, including an algorithm for cognate extraction.

1 Introduction

Yiddish is a Germanic language spoken mainly by Jews of Eastern European (Ashkenazi) descent. The language is currently spoken by perhaps 3 million people (Gordon, 2005) primarily in the US, in Israel and throughout the world. Prior to World War II there were 11-13 million speakers (Jacobs, 2005), and a rich literature in Yiddish was being produced, but the number of Yiddish speakers has since dramatically declined both due to death of many of them in the Holocaust and to post-war assimilation. The number of speakers continues to decline as Yiddish-speaking population ages and the language is listed as *definitely endangered* by UNESCO (2009).

In this paper we describe high-quality machine translation systems for Yiddish-English and English-Yiddish. We are not aware of any machine translation system for these language pairs, although online dictionaries exist. There is existing work on building MT systems for languages with scarce resources, e.g. Nießen and Ney (2004), Wang et al (2006), Al-Onaizan et al (2000). Much of the work of this sort is language-specific by nature, but in general it tends to explore morphological features of the rare language, or it involves translation between closely related languages, like Punjabi to Hindi (Josan and Lehal, 2008), or Czech to Russian (Hajic, 1987). We will use elements of both of these approaches.

2 Challenges

Yiddish has some unique challenges which we needed to overcome to produce a high-quality system.

2.1 Orthography

Yiddish is written in a unique script which is based on Hebrew, but has some important differences. Unlike Hebrew, vowels are written and there is a general correspondence between spelling and pronunciation. However, Yiddish spelling is not entirely standardized and depends on the dialect of the speaker, country of origin, and time of writing. Until the appearance and general use of Unicode, there was no standard way to encode Yiddish text. As a result, large amount of Yiddish data on the web is in the form of rendered PDF files (which would require OCR to decipher) or in transliterated form, using Latin script. A standard of Yiddish transliteration exists, known as YIVO[1]. The standard, however, is not accepted by all writers, who tend to have different preferences for various systems. Furthermore, while the YIVO standard is deterministic in producing transliteration, it is non-deterministic in reverse transliteration, so in general we cannot obtain the original Yiddish from its transliterated form.

Yiddish orthography utilizes combining marks, known as points, to differentiate Hebrew letters. For instance, the letter *beyz* (Hebrew *bet*) can be written without points and be pronounced as **[b]**, or with a *rafe* point which causes it to be pronounced as **[v]**.

[1] http://www.yivoinstitute.org/yiddish/alefbeys_fr.htm

This convention may require two Unicode points to represent a single letter. A number of sources which do not support Unicode properly or desire to avoid complexity, simply drop these distinguishing marks which makes transliteration of names from Yiddish non-deterministic and makes it harder to find cognates in Latin-script German language. To further complicate things, some sources choose to distinguish some letters with points, but not others.

2.2 Data availability

Yiddish possesses a rich religious and secular literature, but much of it is out of print and is unavailable in digital form. The total amount of available digital data is quite low and is increasing very slowly. Most modern Yiddish speakers are literate in another language (local vernacular or Hebrew for religious purposes) and primarily produce written output in that language rather than Yiddish, relegating Yiddish for primarily oral use within their community. There is, however, an ongoing project to preserve Yiddish literature by collecting and scanning out of print books by National Yiddish Book Center and we may benefit from it. We are, however, currently hampered by the lack of high quality OCR for Yiddish.

Some Yiddish data does exist, e.g. Yiddish language Wikipedia and Wiktionary, a Yiddish newspaper *Forverts*[2], various newsletters and blogs, some digitized literature (mostly short stories), and some discussion forums. Very little of this data is parallel. There are online dictionaries, though the major commercial one, by Ectaco[3], is transliterated into Latin script using a non-YIVO (but very similar) standard.

2.3 Language properties

Yiddish is an amalgamation of three different sources: Germanic, Hebrew and Slavic. Its grammar is closely related to that of German, which is also the largest source of its vocabulary. Hebrew contributed most of the religious terms, but also some common words. Slavic influence (primarily Russian and Polish) is the smallest of the three groups. Words which are derived from Hebrew tend to keep Hebrew spelling (i.e., no vowels) and use some letters not used by Germanic and Slavic sources. In modern Yiddish text one also finds a large number of words borrowed from a local language (usually English) adapted to sound Yiddish. These may include neologisms, but also words made up on the spot by the bilingual speaker because he or she does not know the "proper" Yiddish word.

Yiddish has significant morphology. Adjectives decline and verbs conjugate by changing suffixes, but also with Germanic umlaut (changing the vowels in the stem). There is also a number of productive suffixes to form words indicating professions, qualities, etc.

Yiddish word order is similar to that of German (but closer to English) and thus may require non-local reordering (especially for verbs) for translating to/from English.

For an overview of Yiddish grammar and morphology, see Jacobs (2005).

3 Our System

We are using a state of the art phrase-based MT engine, similar to that of Och and Ney (2004). Such a system requires parallel data for training, and also monolingual data for the target language to create a language model.

3.1 Training Data

There is no standard parallel corpus to be used for training. We used some heuristic methods to match URLs and verified them by language detection to find some 200K words of noisy parallel text. Much of it is either not truly parallel, or not entirely in Yiddish, or both. Even if all of it were perfect, this amount of data is not usually sufficient to get reasonable MT quality.

There are some Yiddish-English dictionaries available online. We were able to extract about 13K entries from one of them (Finkl, 2009) available under a free license. A commercial dictionary by Ectaco (which we licensed) contains about 13K entries of transliterated Yiddish and was not directly usable. We have detransliterated it back into Hebrew script by picking most likely source characters that could give rise to the transliteration, obtaining a noisy dictionary.

3.2 Transliteration

Since Yiddish is written in a script that is different from Latin, we find it necessary to transliter-

[2] http://yiddish.forward.com/

[3] http://www.ectaco.com

Letter	Name	Transliteration
אַ	pasekh alef	a
אָ	komets aleph	o
ב	beys	b
פּ	pey	p
פֿ	fey	f
ש	sin	s
ת	sof	s
ס	samekh	s
יַ	pasekh tsvey yudn	ay

Table 1: Sample YIVO transliterations rules

ate in both directions. Names of people and organizations are transcribed either phonetically or according to spelling (or both). YIVO transliteration standard prescribes a certain set of rules for rendering Yiddish words (such as names) into English and we implemented these rules. This is a reasonable last-resort approach for translating unknown Yiddish words into English, so that the translation output is entirely in Latin script. Unfortunately, much of our input does not use *points* to distinguish letters. As a result we end up with incorrect transliterations. Furthermore, most words we need to transliterate from Yiddish are names of non-Yiddish origin, and their transliteration produces similar-sounding, but incorrect forms (e.g., *Klintan* for *Clinton*). YIVO transliteration is also not designed for transliterating English words into Yiddish, but we also use it for English-Yiddish transliteration as a last resort. In some cases, where English spelling is not monotonic sound-to-letter correspondence, we obtain very strange transliterations, e.g. the transliteration of *Lane* would be pronounced something like **[lanɛ]** rather than **[leɪn]**. Some examples of YIVO transliteration rules are given in Table 1.

To solve the problem of phonetic transliteration, we utilize an English pronunciation dictionary CMUDict (Rudnicky, 2009), and map its phonemes into letters of Yiddish alphabet. For each pronunciation entry we thus update a tentative English-Yiddish phonetic transliteration. We use these transliterations as a last resort for both translation directions whenever either English or Yiddish side matches and when no genuine translation is available. This works very well, but introduces spurious words like *Aliens* for *Alliance*. See Tables 2 and 3 for details.

CMU Phoneme	IPA	Yiddish mapping
AY	[aɪ]	יַ
B	[b]	ב
CH	[tʃ]	טש
S	[s]	ס

Table 2: Some CMU phonemes and their mapping to Yiddish

Word	Pronunciation	IPA	Yiddish mapping
Lane	L EY N	[leɪn]	ליין
Oops	UW P S	[ʊps]	וּפּס
Purple	P ER P AH L	[pɜrpɒl]	פּערפּאַל

Table 3: Some CMU dictionary entries and their mapping to Yiddish

3.3 Language Model

It is necessary to build a Yiddish ngram language model to obtain reasonable quality for the English-Yiddish system. For this we need to accurately identify Yiddish text on the web. If we restrict ourselves to those web pages which are predominantly in Hebrew script, we find that 99% or more of this text is in Hebrew. We are thus subject to the false positive paradox: a highly accurate language identifier will still misidentify a large number of Hebrew documents as Yiddish. Hebrew is thus present in our language model training data, although we do not expect it hurts MT quality to a large extent, since our MT system is unlikely to produce Hebrew in its output. We cannot produce a completely Hebrew-free language model to test this hypothesis, but we were able to do the reverse: training a language model with an overwhelmingly large amount of Hebrew added had no impact on MT performance (although, as expected, cross-entropy of this model against any Yiddish text is much higher). This means, however, that the statistics below may be quite inaccurate in terms of the actual amount of Yiddish data we find.

We identified some 300K documents containing 700 million word tokens and 1 million word types as Yiddish, most of these being noise. After filtering we obtain 65 million ngrams (up to length 4), of which 20 million are unique. Probably only a portion of these is genuine Yiddish.

Type	List of items
Adjective suffixes	-er, -n, -e, -es, -en, -em
Verb suffixes	-n, -st, -t
Verb prefix	ge-

Table 4: Affixes used in morphology normalization

The resulting language model requires an important modification. Because *points* are optional they need to be normalized away for the purposes of language modeling, even though our MT system generates words with *points*. This makes our language model unsound, since probabilities of words in a given ngram context could sum up to more than 1. However, this modification seems to improve translation quality in practice.

3.4 Morphology Normalization

We perform morphology normalization on the source side, stripping suffixes and prefixes used in declination and conjugation. Our phrase table entries are then keyed by the normalized forms, although unnormalized forms are also stored. If an exact match for an unnormalized source phrase is found, corresponding entries are returned. If there is no exact match, we return phrases that would match the query after normalization.

A list of affixes used is given in table 4 in YIVO transliteration. These map together different cases and genders of adjectives, and persons and tenses of the verb. The latter may cause the verb to translate in the wrong tense, but only if there is no exact match, so it is better than not translating it at all.

3.5 Spelling Correction

As we discussed above, Yiddish spelling, especially in the case of names, is often somewhat variable. In addition, lack of a convenient way to input Yiddish characters causes a large number of typos. We can use the language model to do spelling correction: when a given word is not present in the phrase table (even after morphology normalization), we try modifying each word by a single insertion, deletion, substitution, or swapping of adjacent letters. If the resulting word exists in our phrase table, and its language model cost in its context is better than that of the original word by a fixed margin, we correct the word in question.

3.6 Cognate Extraction

The measures described above try to address the problem of unknown words; they are, however, insufficient to solve it. We have too little training data to use and the number of out-of-vocabulary words is high enough to prevent understanding what a sentence means. Most of Yiddish vocabulary is German in origin, and most of the remainder derives from Hebrew. This makes it possible to extract dictionaries for Yiddish-German (and Yiddish-Hebrew) from monolingual data.

The algorithm is based on the nearest neighbor algorithm using an adaptive Levenshtein distance (Levenshtein, 1966) where substitution, insertion and deletion costs are computed with maximum likelihood estimation. The details are given in Algorithm 1.

The basic idea is as follows: for every Yiddish word, find the nearest German word. If that word is close enough, do the following: record which operations lie on the optimal comparison path; e.g., for words *zabcd* and *axbce* we record substitutions (a, a), (b, b), (c, c), (d, e), insertion x, and deletion z. These counts are then used to obtain new operation costs as negative log probability estimated using maximum likelihood. Because a substitution can be viewed symmetrically from either side, we take the smaller of the two costs; whereas for insertion or deletion there is only one side and we simply use the MLE from that side. In practice, we smooth all probabilities, to ensure that new characters are allowed to be substituted with high, but finite cost. The process is then repeated.

At the end we obtain a Yiddish-German dictionary. We used 100K word types of Yiddish and the same number for German as input, and generated some 45K word pairs. By looking at a sample of 100 entries, a German speaker estimated about 74% precision. We do not have a good measure of recall. Our precision is higher if fewer words are used as input, but we generated fewer entries overall which lowers recall. By experimentation on development set we found that using 100K words produces the highest BLEU score.

Algorithm 1 Nearest neighbor with adaptive Levenshtein distance

input: Y = Yiddish word list (transliterated)
input: G = German word list
$\forall i, j \quad SubCost(i, j) = 1 - \delta_{ij}$
$\forall i \quad InsCost(i) = DelCost(i) = 1$
$Costs = \{SubCost, InsCost, DelCost\}$
for several iterations **do**
  $\forall i, j \quad SubCount(i, j) = 0$
  $\forall i \quad InsCount(i) = DelCount(i) = 0$
  $Counts = \{SubCount, InsCount, DelCount\}$
  for $\forall y \in Y$ **do**
    $bestG = \arg\min_{g \in G} Levenshtein(Costs, y, g)$
    if $Levenshtein(y, bestG) < threshold$ **then**
      $IncrementCounts(Counts, y, bestG)$
    end if
  end for
  $\forall i \quad TotalSrcCount(i) = DelCount(i) + \sum_j SubCount(i, j)$
  $\forall j \quad TotalTrgCount(j) = InsCount(j) + \sum_i SubCount(i, j)$
  $\forall i, j \quad SubCost(i, j) = -\log \frac{SubCount(i,j)}{\min(TotalSrcCount(i), TotalSrcCount(j))}$
  $\forall i \quad DelCost(i) = -\log \frac{DelCount(i)}{TotalSrcCount(i)}$
  $\forall j \quad InsCost(j) = -\log \frac{InsCount(j)}{TotalTrgCount(j)}$
end for

3.7 Bridging

In the previous section we were able to obtain a reasonable quality Yiddish-German dictionary. Since Hebrew words are written in the same script, they tend to be adapted into Yiddish without modification and it is trivial to match them up and get a Yiddish-Hebrew dictionary with a remarkable precision (over 90%, as estimated by a Hebrew speaker), but low recall. However, we need this data for Yiddish-English. To obtain it, we use phrase table bridging (also known as pivoting) which is commonly used for this purpose (see Habash and Hu (2009) for one recent example).

We train Hebrew-English and German-English systems, which have much better quality and larger vocabulary coverage than our Yiddish-English system. For any phrase in the original phrase table, we substitute each source word by a corresponding Yiddish word using our dictionary. If the entire source phrase can be covered, we output it. If it can be covered in multiple ways, we output target phrases for each of the resulting source phrases. For the German-Yiddish system we augment each resulting phrase by an extra feature which indicates the Levenshtein distance between German and Yiddish words.

We also used a free Polish-Yiddish dictionary[4] and performed bridging over Polish.

4 Results

There is no standard data set to be used for Yiddish-English translation. We created a test set by translating 1000 sentences of WikiNews data into Yiddish (and some other languages). We can thus estimate the quality of our Yiddish-English system as it compares to other X-English systems on the same set. We also tested our English-Yiddish system on the same set, but those numbers are not comparable to anything else, although they can be used to track progress.

BLEU scores (Papineni et al., 2002) are reported for several Germanic languages in Table 5. A paragraph translated from Yiddish to English, along with the same paragraph translated from German and Afrikaans, and the English reference are provided in

[4]Online, at http://pl.wiktionary.org/wiki/Kategoria:Jidysz

Source language	Training data (million words)	BLEU
German	630	23.00
Dutch	520	28.29
Icelandic	21	25.51
Afrikaans	13	18.03
Yiddish	0.2	19.69

Table 5: BLEU scores on the same set for some Germanic languages

Table 6. We provide both an initial version, which uses nothing except automatically mined data, as well as the final version, which uses all the modules described above. It seems that BLEU score is not entirely adequate for comparison across different language pairs, as the sample translation indicates. Despite BLEU scores being quite high compared to other languages, Yiddish-English quality is substantially lower.

For the benefit of Yiddish-speaking readers, we also provide a translation of the English reference into Yiddish by our English-Yiddish system. This output is provided in Yiddish script and in YIVO transliteration.

The lack of training data is quite apparent in the English-Yiddish passage, which has a number of unknown words and spurious translations. Nonetheless, it is possible to understand the passage's meaning.

In Table 7 we list the breakdown of contributions to the final BLEU score for each of the elements of the Yiddish-English system. We can see that mined parallel data alone, while important, gave a relatively small part of the gains, and dictionaries, even though small, helped significantly. By far the most important of the remaining contributions was the use of cognate extraction via German.

Our English-Yiddish system quality is harder to evaluate since it requires a native speaker of Yiddish. We obtain BLEU score of about 10%.

5 Discussion and Future Work

As far as we aware, we have created the first Yiddish-English and English-Yiddish automatic MT systems. This presented us with some unique challenges dealing with data scarcity, complicated orthography and morphology. Our Yiddish-English system quality is quite high for such a small amount of data available.

Item	**Score impact**
Mined parallel data	+6
Dictionaries	+5.5
Morphology	+0.5
Transliteration	+1.5
Bridged (German)	+3.5
Bridged (Hebrew)	+0.5
Bridged (Polish)	+0.5
Spelling correction	+1.5
Total	19.5

Table 7: Yiddish-English contributions

We intend to improve our system further by applying more bridging to take advantage of any other Yiddish-X data that may be available. We expect to obtain significant gains by digitizing Yiddish books via OCR and aligning them to their translations in other languages, using bridging in the cases where no English translation is available.

We will use native speakers of Yiddish to evaluate and help us improve our English-Yiddish system.

We intend to apply lessons learned in this effort to other languages with scarce resources. We consider working on such languages a very important public service that will help preserve these languages and make literature in these languages available to the rest of the world.

Acknowledgements

We would like to acknowledge the contributions of the members of the machine translation team at Google, especially Franz Och, as well as from Hesky Fisher who as a native speaker of Yiddish provided valuable guidance. In addition, we would like to mention the work of National Yiddish Book Center and its founder, Aaron Lansky, whose book *Outwitting History* inspired one of the authors to start this work.

Language	Example passage
Yiddish to English: initial version, only parallel data and YIVO transliteration	lem, farshribn oyi 'one of the Lesin nshumus of the time' was the author of tsendlikher bikher, which is ha Nazis in 41 languages and Aliya more than 27 milyanen kafyes. He has pupils an International shm for the tsiberyad, a Milman Gadya mesius about a mekhanisher world regirt of bots, arusgegebn first in English in 1974. trats be alveltlekher popularity, is lem not a mitgid of science - fiktsye and fantazye spoken of de.
Yiddish to English: current version	People, require them 'one of the profound minds of the time, was the author of finite Books, which eh translations in 41 languages and sold more than 27 million copies. He has gained an international reputation for the tsiberyad, a series of short stories about a mechanical world ruled by robots, first published in English in 1974. Despite his worldly popularity, is people not a member of the science - fiction and fantasy Screw of America.
Afrikaans to English: for comparison	As that described as one of the diepsinnigste minds of his time, has written numerous books, which translates into 41 languages and with more than 27 million copies sold. He has earned international volubilis with the Cyberiad, a series of stories about a world of machines, which are ruled by robots. In 1974 the first published in English. Despite its international popularity was not as a member of SFWA (Science Fiction and Fantasy Writers of America) no.
German to English: for comparison	Lem is considered "one of the most profound minds of our time." His many works have been translated into 41 languages, reaching a total circulation of 27 million copies. He gained international fame with "Kyberiade" - in English first published in 1974 - a series of short stories, in a mechanical, robot-dominated world. Despite its international popularity, Lem was a member of the Science Fiction and Fantasy Writers of America.
English reference	Lem is considered "one of the most profound minds of our time." His many works have been translated into 41 languages, reaching a total circulation of 27 million copies. He gained international fame for The Cyberiad, a series of short stories from a mechanical world ruled by robots, first published in English in 1974. Despite his international popularity, Lem is not a member of the Science Fiction and Fantasy Writers of America.
English to Yiddish: above passage translated to Yiddish	לעם, דעסcריבעד וווי "איינער פֿון די טיף שטימונג פֿון דער צייט," האָט דער מחבר פֿון דאָזענס פֿון באָאָקס, וואָס זיַנען איבערגעזעצט אין 41 שפראַכן און סאָלד איבער 27 מיליאָן cאָפּיעס ער גאַינעד אינטערנאַציאָנאַל רום פֿאַר די cיבעריאַד, א סעריע פֿון קליין סטאָריעס פֿון א מעכאַניש וועלט רולעד לויט ראָבאָץ, ערשט פֿאַרעפֿנטלעכט אין ענגליש אין 1974. טראָץ זיין אינטערנאַציאָנאַל פּאָפּולאַריטעט, לעם איז נישט א מיטגליד פֿון דער וויסנשאַפֿט פיקשאַן און פאַנטאַסי וועריטערס פֿון אַמעריקע.
English to Yiddish: above passage in YIVO transliteration	lem, described vi "eyner fun di tif shtimung fun der tseyt," hot der mkhbr fun dozens fun books, vos zaynen ibergezetst in 41 shprakhn aun sold iber 27 milyon copyes. er gained internatsyonal rum far di ciberyad, a serye fun kleyn storyes fun a mekhanish velt ruled loyt robots, ersht farefntlekht in english in 1974. trots zeyn internatsyonal popularitet, lem iz nisht a mitglid fun der visnshaft fikshan aun fantasi vriters fun amerike.

Table 6: Example passage

References

Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Daniel Marcu, and Kenji Yamada. 2000. Translating with scarce resources. In *Proceedings of the National Conference on Artificial Intelligence (AAAI*, pages 672–678.

Refoyl Finkl. 2009. Yiddish-English dictionary. Online, http://www.tichnut.de/jewish/yiddishdictionary.html.

Raymond G. Gordon, Jr., editor. 2005. *Ethnologue: Languages of the World*. Dallas, Tex.: SIL International, fifteenth edition.

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese statistical machine translation using English as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece, March. Association for Computational Linguistics.

Jan Hajic. 1987. Ruslan - an mt system between closely related languages. In *EACL*, pages 113–117.

Neil G. Jacobs. 2005. *Yiddish: a Linguistic Introduction*. Cambridge University Press, Cambridge.

Gurpreet Singh Josan and Gurpreet Singh Lehal. 2008. A Punjabi to Hindi machine translation system. In *Coling 2008: Companion volume: Demonstrations*, pages 157–160, Manchester, UK, August. Coling 2008 Organizing Committee.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.*, 10(8):707–710, February.

Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL-02*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Alex Rudnicky. 2009. The Carnegie Mellon pronouncing dictionary, version 0.7a. Online.

UNESCO. 2009. UNESCO interactive atlas of the world's languages in danger. Online, at http://www.unesco.org/culture/ich/index.php?pg=00206.

Haifeng Wang, Hua Wu, and Zhanyi Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 874–881, Sydney, Australia, July. Association for Computational Linguistics.