Applying Cross-Entropy Difference for Selecting Parallel Training Data from Publicly Available Sources for Conversational Machine Translation

William D. Lewis, Christian Federmann, Ying Xin

Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
{wilewis,chrife,v-yixi}@microsoft.com

Abstract

Cross Entropy Difference (CED) has proven to be a very effective method for selecting domain-specific data from large corpora of out-of-domain or general domain content. It is used in a number of different scenarios, and is particularly popular in bake-off competitions in which participants have a limited set of resources to draw from, and need to sub-sample the data in such a way as to ensure better results on domain-specific test sets. The underlying algorithm is handy since one can provide a set of in-domain data and, using a language model (LM) trained on this in-domain data, along with one trained on out-of-domain or general domain content, use it to "identify more of the same." Although CED was designed to select domain-specific data, in this work we are generous regarding the notion of "domain". Instead of looking for data of a particular domain, we seek to identify data of a particular *style*, specifically, data that is conversational. Our interest is to train conversational Machine Translation (MT) systems, and boost the available data using CED against large, publicly available general domain corpora. Experimental results on conversational test sets show that CED can greatly benefit machine translation system quality in conversational scenarios, and can be used to significantly increase the amount of parallel conversational data available.

1. Introduction

Cross Entropy Difference (CED) as defined by [1] has proven to be a very effective method for selecting domain-specific data from a larger corpus of out-of-domain or general domain content. It is used in a number of different scenarios, and is particularly popular in bake-off competitions—such as those hosted by the WMT [2] or IWSLT [3]—in which participants have a limited set of resources to draw from, and need to sub-sample the data in such a way as to ensure better results on domain-specific test sets. It has also proven useful in scenarios where training on all available data is not possible or feasible, or where iterating on large samples of data takes too long [4].

The algorithm is handy since one can provide a set of in-domain data and, using an LM built over the in-domain data, use it to "find more of the same" in a larger store of parallel or monolingual data. Although the output generated by CED may not truly be *in-domain*—Axelrod et al 2011 [5] use the term "pseudo in-domain"—the resulting data generally proves useful enough, and quality on relevant, in-domain, test data improves sufficiently enough, to warrant CED's inclusion in one's "bag of tricks" for manipulating data for SMT or language model building.

Although CED was designed to select domain-specific data, in this paper we are generous regarding the notion of "domain". Since we are looking for data not necessarily of a particular domain but rather we are looking for data of a particular *style* or *register*, that is, conversational. People have conversations about just about anything, so conversations truly defy domain.

Our primary interest, however, even more than using CED for style adaptation, is to find a means to bolster the amount parallel conversational data that is available for training conversational MT systems—essentially MT systems that we could be used in an end-to-end speech-to-speech (S2S) pipeline. Conversational data, specifically fluent transcripts of conversations, especially *parallel* conversational data, is very difficult to come by; only a very small set of language pairs have any parallel conversational data, and the quantities that are available are quite small. By contrast, the amount of broad-domain parallel data that is available has grown dramatically over the past few years (*e.g.,* CommonCrawl, EuroParl, United Nations, etc.). Enter CED as a method to find conversational content in the much larger stores of heterogeneous, general domain data.

We assume that a conversational MT system must be able to take as input the transcripts of speech recognition (*a la* [6]). We assume further that we have a mechanism to clean up disfluencies in the source ASR output in order to make it more hospitable to an MT engine (how to do such data cleaning is beyond the scope of this paper;[1] we assume clean input

[1] We employ a method for such data cleaning called *TrueText*. [7] gives some background on how producing "fluent" content from speech recognition can improve downstream processes, such as Machine Translation. Given space limits, we will not expand upon TrueText in this paper, but suggest the reader explore [7] for more background.

for the MT, effectively constituting “oracle” output from the ASR[2]. To this end, we seek to use CED to bolster the amount of *parallel* conversational-style (or “pseudo-conversational-style”) data available to us. Using a method to discover conversational content, notably *parallel* conversational content, can help build more robust conversational MT systems.

To determine the utility of data output by CED for this task, we measure end-to-end MT results on conversational test sets representative of actual mono- and bi-lingual conversations. For our general domain corpora, we draw from all publicly available parallel sources for English↔French that we know of (shown in detail in Table 3). Combined and added to training, these sources act as our general domain source data and our ceiling (when we train on all of the data). To test a “what if” scenario—that is, “what if” we had a much larger store of data available to draw from beyond those that are publicly available—we use CED against a very large store of Web-scraped English↔French content (over 500 million parallel sentences) combined with the publicly available data to create another ceiling. With this ceiling we show that CED can expand to much larger stores of data, and demonstrate the gain others can reasonably expect to see using this method in the near-term. Experimental results on conversational test sets show that style adaptation using CED greatly benefits MT quality in conversational scenarios.

This paper is organized as follows: Section 2 provides more details on the CED method while Section 3 explains our experimental setup and the data we have used. We discuss results in Section 4 and conclude with a summary and an outlook to future research questions in Section 5.

2. Background

2.1. Cross-Entropy Difference

The intent of the Cross Entropy Difference (CED) algorithm [1] is to identify a subset of data in a much larger corpus of data that is in the “domain” of interest. Using an in-domain corpus, and an LM built over the corpus, we can find more content that resembles the domain of interest. The CED algorithm, as shown in Figure 1, relies on three principal components: (i) an in-domain LM S_{in} (or LMs, in the case of [5]), (ii) an out-of-domain LM S_{out}, and (iii) an out-of-domain or general domain corpus from which we are selecting data ((ii) can be built over the data in (iii), but that is not required). For each sentence in (iii) s_i, the CED algorithm calculates the cross entropy from the in-domain LM S_{in}, and subtracts from it the cross entropy for the same sentence scored against the out-of-domain LM S_{out}. Although one would expect scoring against the in-domain should be adequate in and of itself, *e.g.,* one would expect the entropy of sentences that share characteristics of the domain, *e.g.,* shared n-gram frequencies, would be adequately scored against the in-domain LM S_{in}. This is the thinking behind related and earlier attempts at the same [8, 9]. However, by simultaneously scoring against an LM built over content that is not in the domain of interest, we favor content that scores *better* on the in-domain LM and more *poorly* against the out-of-domain LM. This, in effect, “pushes” the selection towards in-domain content and away from out-of-domain content. Figure 1 shows the algorithm.

$$\mathcal{CED}(s_i|S_{\text{in}}, S_{\text{out}}) = H_{\mathcal{LM}(S_{\text{in}})}(s_i) - H_{\mathcal{LM}(S_{\text{out}})}(s_i) \qquad (1)$$

The most common usage of CED in MT, as noted earlier regarding *bake-offs*, has been to find additional content in a particular domain, say “news text”, in an out-of-domain corpus, say “parliamentary proceedings”, *e.g.,* Europarl [10]. We may or may not have bilingual data for the in-domain corpus, but if we do we can pool it with a set of data selected by CED, and use it for training our in-domain translation models. The percentage of content that we should select is often decided upon by trial and error, that is, select 5%, 10%, 15%, etc., of the data desired, and where quality plateaus, select that percentage. Since CED assigns a score to every sentence for an out-of-domain corpus, we can rank the data by that score, and select the top n% from the ranked data, and then train our models on that percentage.

2.2. The Nature of Conversational Data

The definition of what constitutes a domain has mostly been avoided in the MT literature. Researchers will generally refer to a domain by name, *e.g.,* news, blogs, government, tech, etc., without ever really defining what the characteristics of that domain are. For conversational data, which is really not a domain at all but rather a *style* or *register*, *i.e., a manner in which language is used*, we can be a little clearer in our definition. There are a number of features that characterize data in the conversational style, among them being what is shown in Table 1. Given that most of these features can be captured by simple LMs, their presence can be boosted by CED.

3. Data and Experiments

3.1. Data Sources (for Training and Tuning)

In this section we provide detail on the data we use in our experiments:

Publicly available data sets – Table 2 shows the sets of data that are available publicly as well as their sizes. This data serves as our general-domain content (our S_{out}) for the set of experiments against which we apply CED (and we also use it for producing our Ceiling System (D), and we randomly sample it for control baselines (B)).

CED seed data – Our seed, in-domain corpus is drawn from the Fisher Corpus [11], and consists of 760K English sentences. The Fisher Corpus consists of transcripts

[2] See [6] for an alternative approach.

Id	Feature	Description / Examples
F1	Increased use of contracted forms	*don't, can't, I'm, I'll, you're*
F2	Increased use of reduced forms	Forms common in colloquial speech, *e.g., gonna, wanna, shoulda, musta, kinda*
F3	Increased use of slang	
F4	Higher frequency of 1st and 2nd person	1st and 2nd person pronouns and verbal forms are more common in colloquial speech vs. Web content as a whole
F5	Shorter Sentences	Conversational utterances tend to be shorter than many sources of textual content
F6	Reduced vocabulary	
F7	Sentence Fragments/Partial Utterances	
F8	Disfluencies and Restarts	Disfluencies: *um, uh, you know, I mean* Restarts: *I I, I'm uh I've*

Table 1: Features of the conversational style

Source	Sentences	Words (English)
Common Crawl 2015	2.98M	58M
Europarl v7 2015	1.79M	43M
FBIS	38K	851K
Gutenberg (No Shakespeare)	196K	3.1M
JRCDGT	698K	15.8M
JRC	1.87M	45.3M
MultiUN	9.1M	228.6M
Subtitle2012	13.8M	96.8M
Subtitle2013	15.1M	106.6M
WIT3	167K	2.5M
WMT2009 Giga	23.93M	532.8M
WMT2009 News	64.6M	1.33M
WMT2011 News	117K	2.5M
WMT2012 News	139K	2.91M
WMT2013 News Commentary	158K	3.4M
WMT2014 News Commentary 2015	179K	3.8M
Total	70.4M	1.15B

Table 2: Publicly available data comprising our general pool

Data set	Sentences	Words
Baseline (A)	22,912,400	167,690,601
Baseline (B)	7,288,000	167,127,882
Baseline (C)	14,300,000	166,085,537
Ceiling (D)	60,864,815	1,037,969,219
Ceiling (E)	93,700,367	1,145,178,939

Table 3: Overview on training data sets for our experiments

of over 2,000 hours of English-speaking phone calls. These are unscripted and, hence, very conversational.

Training data – Core to one of our baseline systems (A) is just the set of Open Subtitle content. We assume that subtitle data is reasonably conversational (albeit scripted), and thus makes a good "core" set of training data for conversational MT. It acts as our primary baseline. To (A), we add varying amounts of "Style Adapted" (SA) data. Our SA data consists of four different sets, specifically 10%, 20%, 30%, and 40%, ranked by CED, drawn from the publicly available data shown in Table 2.[3] Our Random Baseline system (B) consists of a random sample of our public data, with approximately the same word count as (A). To be a useful control against (A), we again add the 10-40% SA samples. Baseline (C) is a system containing just the 20% SA sample, and nothing else. Its word count is approximately the same as (A) and (B), and thus can be used for comparison purposes. System (D) was trained on *all* publicly available training data, and thus should act as a *ceiling* system, *possibly* reflecting the peak BLEU scores we might expect to achieve. Finally, system (E) is a system consisting of a very large SA sample, paired with OpenSubtitle content at its core (same assumption as (A) as to the underlying value of subtitle content for conversational systems). The data consists of approximately 94M parallel sentences. The SA data for (E) was drawn from a very large corpus of English↔French Web content, plus all publicly available sources, clocking in at greater than 500 million parallel sentences. We were unable to train another ceiling system on all of this data, so the style-adapted system (E) effectively acts as another ceiling system. The sentence and word counts for each baseline system (A), (B), and (C) are shown in Table 3. We also include the sizes of our two ceiling systems, (D) and (E).

Tuning data – Our dev set is based on a random sample of Web content which contains 6,870 sentence pairs and a total of 123,030 English and 132,903 French words, respectively. Based on our experience with this data set, it can be considered *lightly conversational* as it shares some of the characteristics of conversational

[3] In production, we select the sample, *e.g.,* , 10%, 20%, etc., that produces the highest BLEU score for the particular task at hand. See [5] or [12] for further exploration of the methodology.

data. Still, our main tuning target is general domain text so any measurable improvements on our *strictly conversational test sets* will effectively prove that our data selection approach works as desired.

3.2. Test Data

To test the impact of our data selection on resulting SMT systems, we built several test sets. These are listed below. Crucially, since we wanted to measure the performance of SMT systems on true, open-domain conversational content, each of the **Speech** test sets was created from actual Skype calls that were recorded between participants who were either speaking the same language or different languages (in the latter case, drawn from bilingual conversations).

Supporting real-time, open-domain, bilingual conversations is the gold standard for S2S systems. To evaluate a conversational MT system that performs the translation function in such a system, we felt our test sets had match the scenario as close as possible, that is, be representative of open-domain, conversations. To that end, **Speech**$_{EX,1}$ and **Speech**$_{XE}$ consist of transcripts of English→French and French→English bilingual conversations, respectively, which were then translated into the opposing languages. These tests sets are relatively hard, since they consiste of true, real-time bilingual conversations, but they are also representative of our ultimate S2S goal: to support free-form, open-domain, bilingual conversations between *monolingual* speakers.

1. **Speech**$_{EX,1}$ – This test set consists of the transcripts of the English side of bilingual English↔French conversations. Participants were English↔French bilinguals, who were fully conversant in both languages. In each conversation, one of the two consistently spoke English, the other spoke French. The English transcripts were normalized and then translated into French.
2. **Speech**$_{EX,2}$ – This test set consists of the transcripts of the English side of bilingual English↔French conversations conducted by monolingual speakers, mediated by an S2S system, namely Skype Translator.[4] In other words, each participant spoke in their own language, and the S2S system transcribed and translated their spoken content into the other language. The English audio was human transcribed (the test data does not contain ASR output), normalized, and then translated into French.[5]
3. **Speech**$_{XE}$ – This test set consists of the French side of bilingual English↔French conversations. It is effectively the equivalent of Speech$_{EX,1}$, except in this case the French side data was kept and translated into English. All French data has been recorded by French native speakers so it is an accurate representation of conversational French.
4. **Eval2000**$_{EX}$ – Eval2000 [13] is a standard speech test set consisting of transcripts of English phone conversations. We translated a sample of the Eval2000 transcripts into French in order to create this test set.
5. **Social**$_{XE}$ – This test set consists of a sample of French Facebook posts, which were then translated into English. Although not strictly conversational, Facebook posts, as with any other social media, exhibit some of the features one sees in conversational transcripts.
6. **WMT2013** – This test set consists of a sample of standard test set used at the 2013 Workshop on Machine Translation [2]. It acts as a sanity check. It contains content that is really not relevant to the conversational MT style.

3.3. Experimental Setup

In order to measure the effectiveness on translation quality of data selected using CED, we ran a series of experiments drawing from a general domain pool of English↔French data (our S_{out}). All of the data is publicly available, consisting of corpora such as the CommonCrawl, Project Gutenberg, various WMT data sets, UN data, etc., which are broken down in Table 2.[6] In total, this corpus consists of approximately 70M sentence pairs and 1.15B words (English side), before removing duplicates. The in-domain (or “in-style”) data, or seed data (S_{in}), which is constant in these experiments, consists of a 760K sentence sample from the Fisher data set [11]. Fisher consists of transcripts of *unscripted* phone calls, so the data are quite conversational, and very similar to the **Speech** test sets. We also include in our experiments two ceiling systems, trained on the following data: The first is trained on all available publicly available corpora (effectively, all sources shown in Table 2). The second is a *“what if”* system, trained on 94M sentences, including some 24M sentences discovered using CED from a very large scrape of the Web, consisting of over 500M sentence pairs, which is then combined with other conversational content. The intent of the second ceiling is to demonstrate the potential of CED on very large corpora, and to provide a proof of concept of what is possible as more data becomes publicly available (*e.g.,* as the CommonCrawl data continues to grow). The hypothesis is that as more data becomes available, there will be more snippets of conversational data in the general pool, which increases the amount of beneficial data we extract when we run CED. This in turn will benefit those who are building conversational S2S and MT systems.[7]

[4] Skype Translator is available at the following URL: http://www.skype.com/en/translator-preview/. The functionality of Skype Translator is also being integrated into other Skype versions.

[5] We assume **Speech**$_{EX,2}$ is easier than **Speech**$_{EX,1}$, since users were bound by the current state of the art of the S2S at the time the recordings were made.

[6] Much of this data, specifically, the Europarl data, the CommonCrawl parallel data, and any data sets labeled with “WMT” are available from WMT 2015 [14] at `http://statmt.org/wmt15/translation-task.html`. WIT3 comes from IWSLT [15].

[7] Crucially, CED can be run on corpora of any size. Realistically, the only limiting factors are disk space, the amount of time to run the algorithm over

Our basic experimental setup compares a baseline MT system trained on subtitle data (A) to a contrastive system trained on a set of randomly selected general domain data, basically parallel text harvested from the Web, of approximately the same word count (B). We assume (A) to be conversational (albeit, scripted conversations). To each baseline, we incrementally add samples of style-adapted data, generated using CED from S_{out}. We have an additional baseline of just style-adapted data of similar sizes to (A) and (B), which is composed of just style-adapted content (C). (C) provides a baseline that demonstrates what is possible in conversational MT just using CED (and is size-controlled, having roughly the same sentence count as (A) and (B), and thus directly comparable to these systems). Finally, we train a system on all available general domain data, to act as a "ceiling" (D). All systems are compared against multiple conversationally oriented test data, with a sanity check test set from the WMT, specifically a test set sampled from the WMT 2013 English↔French test data [2].

We use custom tree-to-string (T2S) systems for training the models for our engines. We require a source-side parser for our T2S decoder, which we have for both English and French; for the English→French direction, we use the English parser, and for the opposing direction, the French one. 5-gram Language Models (LMs) are trained over the target-side data for each system. We use Minimum Error Rate Training (MERT) [17] for tuning the lambda values for all systems, and we report results in terms of BLEU score [18] on lowercased output with tokenized punctuation.

4. Evaluation and Analysis

4.1. Experimental Results

Looking at Tables 4 and 5, it is fairly clear that trainings performed on conversational training data fare well on test sets that are conversational in nature. This should not come as a surprise. However, there are some surprises in the results. For the English→French trainings, baseline (C) which consists of just the style-adapted data, outperformed *all* other trainings on the EX Speech-related test sets (having scores of 52.39, 47.39, and 35.45 for $\textit{Speech}_{EX,1}$, $\textit{Speech}_{EX,2}$, and *Eval2000*, respectively), even besting systems trained with subtitle data, including those trained with additional CED "style-adapted" (SA) data (best in class for each test EX set: 51.68 with 20% SA data, 46.59 with 40% SA data, and 34.31 with 10% SA data). What was most startling, however, was that the Random baseline (B) bested the subtitle Baseline (A) on all EX test sets, scoring 50.79, 45.09 and 35.23 versus 50.28, 44.63 and 32.77. This suggests that the subtitle data, contrary to our initial assumptions, is not a good baseline for a conversational MT system. Further, adding SA data for the Random baselines did sometimes improve scores on EX test sets, but no Random baseline+SA pairing bested Baseline (C)—that is, SA data *alone* beats any random baseline—on EX test sets.

Results for the English→French trainings on the XE test sets paint a different story, however. As noted in Section 4.2, there are two XE test sets, $\textit{Speech}_{XE}$ and $\textit{Social}_{XE}$. The former consists of the French side of English→French conversations, and the latter consists of French Facebook posts. Both sets of data were translated into English. On the $\textit{Speech}_{XE}$ test set, the subtitle Baseline (A) beats *all* other results (excepting Big Data (E)), including SA (D) and any combination of SA with Subtitle (A) or Random (B); the baseline score of 51.61 is beat by no training other the Big Data (E). So, contrary to the assessment that subtitle data makes a poor baseline system, it actually proves to be very good *when the data is French sourced*. In fact, it proves to be a much better baseline than SA data (Baseline (C)), completely the opposite of what we saw on the EX test sets. (We examine what the source of this "directionality bias" might be in more detail in Section 4.3.2). On the $\textit{Social}_{XE}$ test set, the SA baseline (C) does equally poorly, beating only the random baseline (23.45 vs. 22.76). Again, since the $\textit{Social}_{XE}$ is French sourced, it provides further evidence of some sort of directionality bias.

For the French→English trainings, the subtitle baseline trainings (A) fare much better than the equivalent EX trainings: on all conversational test sets, they best the SA baseline (C), in some cases paired with varying quantities of SA data. The only odd result is the performance of the Random baseline (B) when paired with 30% SA for $\textit{Speech}_{EX,2}$, which does the best of any system outside of (E). Baseline (B) does very poorly by itself on all test sets, however, performing better when paired with the SA data. SA data, thus, proves to be a useful augmentation for the random baseline (B). SA proves less useful for the subtitle baseline (A) on the EX speech test sets, but much better for the XE speech test set (and the social media XE test set as well). Again, there is evidence here for some sort of directionality bias.

Overall, the SA data contributes. By itself, in the EX trainings, it has proven essential. For XE, it's a useful addition to subtitle data when measured against XE test sets.

4.2. Overview of Experimental Results

Subtitle data appears less useful, but only when either (a) English sourced data is used or (b) training English→French systems. In all other cases, subtitle data proves useful for training conversational MT systems. Domain adapted data, however, proves highly useful for training conversational MT systems. Using existing and readily available public sources of English↔French data, and using existing and readily available monolingual, conversational English seed data, we are effectively able to select "conversational" data from these sources in order to train conversational MT systems with higher BLEU scores. Although SA data has proven universally useful, its value differs depending on the direction of training or test data. In the next section, we examine some

the data—LM scores do not have to be stored in memory, but can be output directly—and building the out-of-domain or general domain LM. Using KenLM [16] for the latter makes CED feasible in most scenarios.

English→French							
Experiment		Test sets					
Data	System	$\text{Speech}_{EX,1}$	$\text{Speech}_{EX,2}$	Speech_{XE}	Eval2000_{EX}	Social_{XE}	WMT13
Baseline (A)	OpenSubtitle	50.28	44.63	**51.61**	32.77	25.27	28.87
+10% SA	OpenSubtitle	51.37	45.36	51.59	*35.02*	25.46	30.80
+20% SA	OpenSubtitle	*51.68*	46.46	51.39	33.95	25.87	**31.43**
+30% SA	OpenSubtitle	51.49	46.54	51.35	33.60	25.75	31.38
+40% SA	OpenSubtitle	51.35	*46.59*	51.07	33.74	**26.10**	31.39
Baseline (B)	Random	50.79	45.09	46.59	*35.23*	22.76	30.39
+10% SA	Random	51.23	46.13	49.68	33.22	24.18	30.50
+20% SA	Random	50.90	45.51	51.07	34.18	**26.10**	30.86
+30% SA	Random	*51.74*	*46.53*	*51.18*	33.87	25.33	*30.97*
+40% SA	Random	51.19	46.19	50.72	33.38	25.40	30.96
Baseline (C)	SA Only	**52.39**	**47.39**	46.45	**35.45**	23.45	30.40
Ceiling (D)	All	50.55	45.86	50.47	32.98	25.67	31.22
Big Data (E)	S2S	58.32	54.04	52.87	37.23	26.65	32.72

Table 4: Translation quality measured using BLEU scores for language pair English→French. Best scores per experiment *in italics*, globally best scores **in bold face**. Table compares Baseline system trained on General domain data to *pseudo in-domain* DomainAdapt system trained on data obtained using the CED method.

distributional clues as to why SA data is useful, and what may be causing this directional discrepancy. The next section constitutes a very preliminary analysis of some of the data and some of the features. We intend to expand this work in the future. What is clear, however, is that there is some sort of directionality bias, and that this bias interacts with the sources of the data.

4.3. A Quick Look at the Conversational Style Features in CED Output

In this section, we look at two main issues: First of all, we look at the distribution of some of the values for a subset of the conversational features, as described in Table 1, across our subtitle, style-adapted, and random baselines, as well as the Fisher corpus we used as our seed data. Second, we compare the distribution of examples of these features in French as well, to see if there are potential discrepancies. We then propose a hypothesis of what might be causing the directionality bias.

4.3.1. Distribution of Conversational Features

In Table 6 we look at the distribution of a subset of the features described in Table 1, specifically, Contractions (F1), Reductions (F2), and 1st and 2nd person forms (F4) (these too are contractions, thus overlap with F1). A comparison between the Subtitle, SA 20%, and the General (Random Sample) shows some interesting tendencies. All three are controlled such that their word counts are roughly the same; the counts in Table 6 are thus effectively normalized (the Fisher data stands out in this regard since it is smaller, and thus is effectively not normalized). Contracted forms, Reductions, and the Distribution of 1st and 2nd person forms are much more frequent in the Subtitle data, suggesting that, if these values are true indicators of conversational content, it is far more conversational. The SA 20% data set is not quite as strong as Subtitle in these feature sets, but it is much stronger than the General data set in both Contracted and 1st and 2nd person forms. Since both SA 20% and the General data were sampled from the same General pool, this provides strong evidence that the CED algorithm, drawing from distributional clues in the Fisher seed data, is selecting a better sample of data for the conversational setting than a random sample does.[8] Noticeably weak in the SA 20% sample are reduced forms, suggesting that they do not occur frequently in the general domain pool (and thus are not available for CED to discover). Thus, in summary, as long as we accept that the distribution of feature values listed here are representative of conversational content, subtitle data does appear to be highly conversational, in comparison with the other data, with the SA 20% data coming in second. These data, in and of themselves, however, do not explain the directionality bias.

4.3.2. The Directionality Bias

We observed in Section 4.1 that our English→French baseline (A) trainings do poorly on English-sourced test data as

[8] It would appear that the LM is, in fact, boosting conversational content based on scoring against the Fisher LM, boosted further by CED due to the absence of these values in the general pool (since those scores are subtracted from the former by CED).

French→English							
Experiment		Test sets					
Data	System	Speech$_{EX,1}$	Speech$_{EX,2}$	Speech$_{XE}$	Eval2000$_{EX}$	Social$_{XE}$	WMT13
Baseline (A)	OpenSubtitle	**55.04**	48.49	51.84	**36.57**	26.77	29.43
+10% SA	OpenSubtitle	54.70	*48.70*	52.60	36.34	27.24	31.61
+20% SA	OpenSubtitle	54.64	48.30	**53.54**	36.22	*27.43*	31.97
+30% SA	OpenSubtitle	53.56	47.61	52.93	35.62	26.89	32.32
+40% SA	OpenSubtitle	53.29	47.67	52.56	35.61	27.36	*32.45*
Baseline (B)	Random	48.33	43.12	46.52	31.81	23.63	31.39
+10% SA	Random	54.11	48.28	*52.78*	35.32	26.59	32.01
+20% SA	Random	53.36	48.39	52.70	35.58	27.07	32.06
+30% SA	Random	*54.39*	**48.79**	52.54	*35.91*	*27.39*	32.27
+40% SA	Random	54.06	48.19	52.64	35.40	27.25	*32.31*
Baseline (C)	SA Only	49.44	44.05	49.37	32.35	23.85	31.48
Ceiling (D)	All	53.73	47.40	52.88	35.38	**27.73**	**32.54**
Big Data (E)	S2S	57.80	51.71	55.54	37.25	27.32	33.30

Table 5: Translation quality measured using BLEU scores for language pair French→English.

compared to our SA baseline (C), but trump baseline (C) for test sets that are French-sourced. Further, we observed that baseline (A) does well on all conversational test sets irrespective of sourcing for the French→English trainings; the baseline (A) trainings beat the SA baseline (C) in all cases. Only on the French-sourced Facebook test set, *Social*$_{XE}$, does baseline (C) show weaker results.

These puzzling results *could* be caused by the discrepancy in conversational features between the English and French sides of our training data. Although we will not find analogous contracted forms in the French, *e.g.,* , for the same person, verbal forms, etc., we can look at the distribution of values for similar features between the two languages. In Table 7 we show values for a small set of French features, namely, (F1) Contractions and (F3) Slang, and a small set of values for each. The (F1) feature is comparable to the same in English in Table 6; (F3) was not tabulated for English, but since the French *argot* forms are often reductions, they are somewhat comparable to (F2) Reductions. When we compare the two tables, Table 7 and Table 6, we can see a much clearer difference between the conversational data (whether seed, subtitle, or SA) and the general data: the ratio of conversational features between conversational vs. general is much larger in French than in English. There are at least two possible reasons for this: (1) English speech is far more colloquial than French, indicated by a higher number of colloquial expressions that occur in conversational data than in written content. Or (2), transcribed English is more likely to preserve the colloquialisms than is transcribed French. (2) could result either from difference in transcription rules between the two languages, or an unconscious bias by French transcribers to avoid transcribing colloquialisms, at least, to avoid transcribing them literally or phonetically.

How might that affect BLEU scores and contribute to a directionality bias? If the English side has a larger number of colloquial expressions, there may likewise be a larger ratio of many-to-one mappings between English and French than in the other direction. In other words, for any given French expression, there will be a higher likelihood of at least two mappings on the English side for that expression (with all the English expressions essentially meaning the same thing, just written differently). Take, for example, the English future marker *gonna*. In formal English, *gonna* is always written as *going to*. A speaker, referring to himself, might say *I'm gonna*, but would never write it that way—*I'm going to* would be the way to write it formally. However, a transcriber, wishing to be true to the input, especially, it would appear, when tasked with captioning movie content, is more likely to write *I'm gonna*. The most common French expression for either is *je vais*, which is the standard form; there is no formal/informal dichotomy for this term in French. In the English→French trainings, both *I'm gonna* and *I'm going to* would resolve to *je vais*, effectively creating a 2:1 mapping, which would have little or no consequence in evaluations on *conversational* test data for the English→French direction. However, in the reverse direction, the 1:2 mapping could lead to occurrences of both forms in the output, causing a failure to match against the test data in a certain percentage of cases, effectively causing a reduction in BLEU scores. Multiplying this effect across the multitude of conversational forms showing in English, and absent in French, could explain the discrepancies observed in the two different directions of the trainings against the test data.

Feature	Seed (Fisher)	Subtitle (A)	SA 20 (C)	General (B)
F1 – Contractions				
don't	81,997	412,479	31,797	9,846
can't	13,717	135,393	9,401	2,707
shouldn't	1,345	13,009	1,156	325
wouldn't	6,439	36,347	2,448	586
couldn't	3,616	26,697	2,767	713
they'll	2,925	2	1,221	295
he'll	1,529	11	535	144
she'll	591	3	161	57
they're	30,713	71,153	7,716	1,780
she's	6,778	77,729	1,452	395
he's	18,842	235,203	4,622	1,352
F2 – Reductions				
gonna	9,588	3,473	4	5
wanna	3,819	960	21	30
shoulda	1	27	2	1
coulda	29	36	2	5
woulda	1	35	1	1
musta	33	1,404	352	707
kinda	7,575	3,671	149	50
F4 – First/Second				
I'm	67,814	460,910	17,981	4,542
I'll	5,735	107	3,775	894
you're	23,699	288,031	13,722	3,508
you'll	1,375	80	7,626	2,232
we're	14,028	125,116	12,589	3,491
we'll	1,817	10	4,000	1,124

Table 6: Distribution of conversational features across different data sets (English-only)

5. Conclusion and Future Work

Overall, the CED algorithm performs well in selecting conversational data from a general pool, as evidenced by the results in both Tables 4 and 5. The algorithm appears to select data in the conversational style, preserving many of the features observed in the conversational source data in the sampled output. The distribution of conversational features in "style" adapted data is not as strong as for conversational data, such as subtitle data, but it still captures a larger sample of conversational features than an equivalently sized random sample does. As shown in the experimentation, "style-adapted" data, that is, data selected by CED, is conversational enough to boost the quality of conversational MT systems. Further, we show that given much larger stores of data, we see even more marked improvements. The continued expansion of the CommonCrawl parallel data, as well as other publicly available sources, can only benefit the larger S2S community as it will consequently increase the pool of readily available (pseudo-)conversational content.

Although we touched upon the directionality bias observed between the English→French vs. the French→English trainings, and hypothesized a potential transcription "bias" between the two languages, the evidence presented was not particularly strong. Since further experimentation with a much larger general pool of data, upwards of 500 million sentence pairs, is showing the same directionality bias effects[9], further investigation in reasons behind this bias is warranted. In our future work, we plan to continue investigating the bias, which includes the exploration of conversational style adaptation for additional languages. We also plan to look at a much more complete set of conversational features (as discussed in [7]). We are also now experimenting with applying CED using other seed sources of data, including data sampled from conversations of Skype Translator users.

References

[1] R. C. Moore and W. D. Lewis, "Intelligent Selection of Language Model Training Data," in *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, Sweden, July 2010. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=138756

[9] These experiments were not included in this paper.

Feature	Subtitle (A)	SA 20 (C)	General (B)
F1 – Contractions			
J'sais pas	171	4	1
T'sais	71	2	1
T'es	52,493	173	76
J'suis	1,071	17	6
M'en fiche	1,185	8	4
F3 – Slang (Argot)			
putain	22,307	72	31
merde	25,486	219	96
ma pote	19	2	1
mon pote	5,234	40	8
meuf	807	13	18

Table 7: Distribution of conversational features across different data sets (French-only)

[2] O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2013 Workshop on Statistical Machine Translation," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 1–44. [Online]. Available: http://www.aclweb.org/anthology/W13-2201

[3] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014," in *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT), Lake Tahoe, CA*, Lake Tahoe, CA, December 2014, pp. 2–17. [Online]. Available: http://www.mt-archive.info/10/IWSLT-2014-Cettolo.pdf

[4] W. D. Lewis and S. Eetemadi, "Dramatically Reducing Training Data Size through Vocabulary Saturation," in *Proceedings of the Eighth WMT*, Sofia, Bulgaria, August 2013.

[5] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of EMNLP*, 2011, pp. 355–362. [Online]. Available: http://research.microsoft.com/en-us/um/people/jfgao/paper/2011-emnlp-camera-select-train-data.pdf

[6] G. Kumar, M. Post, D. Povey, and S. Khudanpur, "Some insights from translating conversational telephone speech," in *Proceedings of ICASSP*, Florence, Italy, May 2014. [Online]. Available: http://cs.jhu.edu/~gkumar/papers/kumar2014some.pdf

[7] E. Fitzgerald, *Reconstructing Spontaneous Speech*. Baltimore, Maryland: The Johns Hopkins University, 2009.

[8] J. Gao, J. Goodman, M. Li, and K.-F. Lee, "Toward a unified approach to statistical language modeling for chinese," in *ACM Transactions on Asian Language Information Processing*, 2002, pp. 3–33.

[9] D. Klakow, "Selecting articles from the language model training corpus," in *ICASSP 2000*, Istanbul, Turkey, 2000, pp. 1695–1698.

[10] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *MT Summit X: Proceedings of the Tenth Machine Translation Summit*, ser. MT Summit '05. Phuket, Thailand: Asia-Pacific Association for Machine Translation, 2005, pp. 79–86. [Online]. Available: http://mt-archive.info/MTS-2005-Koehn.pdf

[11] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: a resource for the next generations of speech-to-text," in *Proceedings of 4th International Conference on Language Resources and Evaluation, LREC*, 2004, pp. 69–71. [Online]. Available: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2004-fisher-corpus.pdf

[12] A. Axelrod, Q. Li, and W. Lewis, "Applications of Data Selection via Cross-Entropy Difference for Real-World Statistical Machine Translation,," in *Proceedings of the IWSLT 2012*, Hong Kong, China, December 2012.

[13] M. Przybocki and A. Martin, "2000 NIST Speaker Recognition Evaluation LDC2001S97," Web Download. Philadelphia: Linguistic Data Consortium, 2001. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2001S97

[14] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi, "Findings of the 2015 workshop on statistical machine translation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1–46. [Online]. Available: http://aclweb.org/anthology/W15-3001

[15] M. Cettolo, C. Girardi, and M. Federico, "WIT3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of EAMT*, Trento, Italy, 2012, pp. 261–268. [Online]. Available: http://hltshare.fbk.eu/EAMT2012/html/Papers/59.pdf

[16] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696. [Online]. Available: http://kheafield.com/professional/edinburgh/estimate_paper.pdf

[17] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st ACL*, Sapporo, Japan, 2003.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th ACL*, Philadelphia, PA, 2002.