Assessing Divergence Measures for Automated Document Routing in an Adaptive MT System

Claire Jaja *+, Douglas M. Briesch *, Jamal Laoudi *+, Clare R. Voss *
*Computational & Information Sciences Directorate, Army Research Lab (ARL), Adelphi, MD
+Advanced Resources Technologies Inc. (ARTI), Alexandria, VA

Abstract

Custom machine translation (MT) engines systematically outperform general-domain MT engines when translating within the relevant custom domain. This paper investigates the use of the Jensen-Shannon divergence measure for automatically routing new documents within a translation system with multiple MT engines to the appropriate custom MT engine in order to obtain the best translation. Three distinct domains are compared, and the impact of the language, size, and preprocessing of the documents on the Jensen-Shannon score is addressed. Six test datasets are then compared to the three known-domain corpora to predict which of the three custom MT engines they would be routed to at runtime given their Jensen-Shannon scores. The results are promising for incorporating this divergence measure into a translation workflow.

Keywords: domain-specific machine translation, divergence measures, document classification

1. Introduction

The advent of robust open-source software for rapidly building machine translation (MT) engines[1] has brought a new challenge to developing translation systems: how to automatically route input texts to the "best" of multiple MT engines available at run-time, given that custom MT engines can systematically outperform general-domain MT engines when translating from the relevant custom domain. For example, *Table 1* shows the accuracy of three translation workflows (rows) on three distinct collections of documents (columns). As can be seen on the diagonal where the *training* domain of the translation workflow matches the domain of the *test* corpus, the accuracy is highest for that workflow (row). Off the diagonal, where the MT training domain and input domain differ, the accuracy drops off dramatically. Our in-house MT engine built with roughly 40k parallel Arabic-English segments from instructional manuals (IM) far outperforms our best available, commercial-off-the-shelf (COTS) engine on that collection. Similarly, a statistical post-editor (SPE) built to run on the COTS output, with only about 1000 segments and 20k named entities from government records (GR), vastly improves the BLEU score for data within the same domain.

The challenge in designing an MT system more generally then is to assess input texts at run-time for routing to the "best-available" MT. With an adaptive workflow, multiple custom and commercial MT engines are available within a single system, allowing for the best possible translation, provided that texts can be accurately routed.

Our approach has been to identify a similarity/divergence measure as a simple predictor for this routing task, such that when given pairs of "unknown-domain" input texts and "known-domain" training corpora, (i) it adequately distinguishes in-domain from cross-domain comparisons, and (ii) the closer its similarity scores, the higher the resulting selected MT accuracy on the input. The Jensen-Shannon (JS) measure met criteria (ii), as shown in *Figure 1*, where the lower (more similar) scores on x-axis, the higher the translation accuracy on the y-axis. With this initial promising result, we focused in detail on criteria (i), assessing the feasibility of using the Jensen-Shannon measure across multiple datasets at runtime to quickly and accurately distinguish in-domain from cross-domain comparisons, in support of an adaptive translation workflow.

	Input Test Domain		
MT by Training Domain	**Broadcast News (BN)**	**Government Records (GR)**	**Instructional Manuals (IM)**
COTS MT General/news domain	**.1775**	.0998	.0702
COTS MT + Custom-built SPE trained on GR data	.1181	**.2133**	.0487
Custom-built MT trained on IM data	.0647	.0511	**.2438**

Table 1: MT accuracy (BLEU score, n4r1) on three test sets run through three different MT workflows

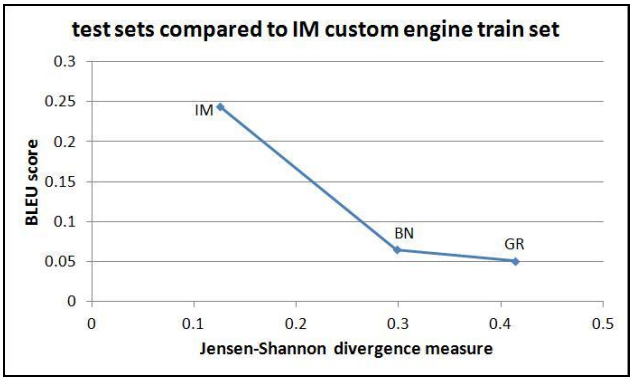


Figure 1: As the test set becomes more divergent from the training set, the MT accuracy (BLEU score) goes down.

[1] Such as MOSES, JOSHUA, and others documented in Sánchez-Martínez and Forcada (2011).

2. Background

Much prior research has been done on domain adaptation for MT and parsing. Research by Gildea (2001) shows that, for parsing, adding a small amount of in-domain training data proves more useful than adding a large amount of out-of-domain training data. Similar results have been shown for MT by Xu et al. (2007) and Koehn and Schroeder (2007) where significant improvement is shown with domain-dependent translation over domain-independent translation.

Many different strategies have been attempted for automatically identifying domain, routing documents, and predicting NLP tool performance.

In parsing, Ravi et al. (2008) use certain characteristics of domains of interest, including sentence length, unknown words, information gain score, and features of the output parse, to accurately predict parser performance. Building on this work, McClosky et al. (2010) train six different language models using six distinct domains and then create linear combinations of these language models at runtime based on the cosine similarity of the 50 most frequent words, the unknown words, and the entropy of the training sets and test sets.

In MT, Zhao et al. (2004) explore unsupervised language model adaptation techniques by using MT output to extract similar sentences and build a domain-specific language model which they interpolate with a general background language model. Banerjee et al. (2010) combine multiple domain-trained translation models using a statistical classifier to classify sentences according to domain. Foster et al. (2010) weight out-of-domain parallel phrase pairs by relevance to the target domain using features such as the number of tokens in the phrase pair, word frequencies, perplexities, unknown words, and an SVM classifier.

Much of this domain adaptation work, however, operates under the assumption that existing corpora are each a domain, and these "domains" are used without internal examination or comparison. Our work focuses on corpus-internal as well as corpus-external comparisons and more generally aims to determine what exactly constitutes a "domain."

Divergence measures, originating from information theory, compare the probability distributions of the elements in two sets. As applied to natural language, this typically includes treating each corpus as a bag of words. Others have previously explored the use of divergence measures for domain identification, document routing, and NLP tool performance prediction.

Lee (2001) estimates word co-occurrence probabilities based on the frequencies of similar co-occurrences. She tests multiple divergence measures, including Kullback-Leibler, Jensen-Shannon, skew divergence, Euclidean, cosine, variational, confusion, and tau. Her best results are using the skew divergence. Pinto et al. (2007) use four different symmetric measures (including the Jensen-Shannon measure), all derived from the Kullback-Liebler divergence, to automatically cluster domain-specific abstracts from scientific papers and technical reports. They find that all four measures perform similarly with results comparable to previous research using the Jaccard similarity measure. Van Asch and Daelemans (2010) explore a number of metrics, including Renyi, variational, Euclidean, cosine, Kullback-Leibler, and Bhattacharyya, to predict the cross-domain performance of a PoS tagger. They show a strong linear correlation between the Renyi measure and the performance loss. Plank and van Noord (2011) use relative word frequencies and topic models paired with similarity functions to automatically acquire related training data in order to parse a given test set. They achieve the best performance with the Jensen-Shannon and variational measures.

Based on this previous research as well as our in-house research applying divergence measures to natural language data, we have found the Jensen-Shannon (JS) divergence measure to be useful in distinguishing between in-domain and out-of-domain data. This measure is a symmetrized and smoothed version of the Kullback-Leibler (KL) divergence which is calculated by the following formula (where *P* and *Q* are the datasets being compared, *p(i)* is the probability of a token *i* in dataset *P*, and *q(i)* is the probability of a token *i* in dataset *Q*):

$$KL(P;Q) = \sum_{i=1}^{N}\left[p(i)\log_2 \frac{p(i)}{q(i)} \right]$$

In the JS formula shown below, as described by Lin (1991), the KL is calculated to a midpoint *M* which equals ½(*P*+*Q*).

$$JS(P;Q) = \frac{1}{2}KL(P;M) + \frac{1}{2}KL(Q;M)$$

This measure has proven useful in part because it is bounded (ranging from 0, indicating identical sets, to 1, indicating non-overlapping sets) and symmetric (meaning comparing Set A to Set B will give the same result as comparing Set B to Set A), as well as consistent, with are measurably lower scores when comparing same or similar domain collections than when comparing different domain collections. Using the Jensen-Shannon measure in an adaptive workflow to automatically route new documents to the MT engine that will provide the best translation seems like a natural step.

Other research has not thoroughly explored the impact of language, size, and preprocessing on this measure. As these factors could all potentially have a significant impact on the measure, we set out to determine how these affect the Jensen-Shannon score.

3. Approach

3.1 Datasets

To begin, we built a range of datasets while varying three properties – language, size, and preprocessing. Our main interest is Arabic-English MT, with a secondary interest in English-Arabic MT, so all of our datasets are in both of these languages. With the exception of one dataset, they are also all parallel datasets.

Given our MT users' interest in low-resource challenges, we also wanted to determine what impact, if any, the size of the datasets being compared has on the Jensen-Shannon measure. We measured corpus size by

the number of tokens, as segments and documents can vary significantly in length, and we calculated the Jensen-Shannon measure on a per-token basis.

In terms of preprocessing, we wanted to test the impact of removing the punctuation, as the punctuation does not carry topic content but might indirectly distinguish genres of the corpora.

All of the English datasets were lowercased, and punctuation was tokenized using NLTK (Bird et al. 2009). All of the Arabic datasets were transliterated using the Buckwalter transliteration schema and morphologically analyzed using ARAGEN (Habash 2004), splitting off all of the punctuation and clitics.

We built three large datasets in different domains. Using the English Gigaword (Fifth Edition) (Parker et al. 2011) and Arabic Gigaword (Fourth Edition) (Parker et al. 2009), both available through the Linguistic Data Consortium (LDC), we built a newswire (NW) domain totaling 1.5 million words. This is the only dataset we use where the English and Arabic are not parallel. Our second domain is instructional manuals (IM) totaling 500k words, originally written in English and translated into Arabic. Our third large domain corpus is government records (GR) totaling 200k words, transcribed from Arabic images and translated into English.

Each of these large datasets are partitioned into a "P" set and a "Q" set to allow for in-domain as well as cross-domain comparisons. From each "P" and "Q" set, we constructed multiple partitioned subsets of 10k words, 50k words, 100k words, 250k words, and 500k words. We also removed punctuation from the original datasets and repeated the same process, to construct additional "P2" and "Q2" sets.

In addition to these large datasets, we built six 10k word test sets in each language, with and without punctuation. Three of these are unused data from the previously mentioned datasets (NW, IM, and GR), and three are completely new domains. The new domains include broadcast news (BN) data from SCOLA, an online language-learning website with transcribed Arabic broadcast news and English translations; earlier government records (earlyGR), transcribed from Arabic images and translated into English; and a Federal Emergency Management Agency (FEMA) document about disaster assistance, originally written in English and translated into Arabic.

3.2 Divergence Measures Tool (DMT)

To provide quick and accurate calculations of the Jensen-Shannon measure, we constructed an in-house software tool (Jaja et al., forthcoming). The tool has a "basic" mode where the user can upload two corpora and see the resulting Jensen-Shannon divergence measure (as well as other divergence measures) and the type and token counts. This mode can display a frequency-sorted or an alphabetically-sorted list of each corpora's types as well as a list of the unique and intersecting types between the two. Additionally, the tool has a "batch" mode where the user can upload multiple datasets, to calculate the divergence measure scores for all pairwise comparisons and then average over these.

3.3 Analysis Phase

For the analysis phase, we report on results with the three large corpora used to calculate within-domain and cross-domain JS similarity scores. The within-domain results are summarized and displayed in matrices with averaged JS scores for all the P subsets of a particular size compared to all the Q subsets of a particular size, for all the available subset sizes for each. The cross-domain result matrices similarly display averaged JS scores for all the P and Q subsets of a particular size for one corpus compared to all the P and Q subsets of a particular size for another corpus, for all the possible pairwise corpora and size combinations. These matrices yield a range of JS scores that can be compared to each other, to determine the impact of text languages (Arabic, English), sizes (10k, 50k, 100k, 500k, 1m tokens[2]), and preprocessing (tokenized with punctuation, tokenized with punctuation removed) on the feasibility of using JS to distinguish in-domain and cross-domain comparisons.

3.4 Evaluation Phase

For the evaluation phase, we report on test results with six test datasets compared to the three known-domain corpora. For the comparison most relevant to our runtime application, we compare 10k token test sets to all the P and Q subsets of each size for each domain.

Our hypothesis for these test datasets is that the three sets built from the same datasets as the large analysis corpora will pattern with their respective domains. Additionally, we expect the BN test set to look most similar to the NW domain, while earlyGR should look most similar to the GR domain. For the FEMA test set, taken from the freely available US "dot gov" websites with parallel Arabic and English documents (among many other languages), we had no a priori sense of which known-domain it would be most similar to.

4. Results and Analyses

4.1 Analysis Phase

The in-domain and cross-domain matrices (see *Figures 2, 3, 4,* and *5*, located at the end of this paper) on the three large corpora provide compelling results for the impacts of language, size, and preprocessing, as well as giving a range for the Jensen-Shannon scores of in-domain versus out-of-domain data. The in-domain results also give insight on the homogeneity of the domains; GR present as the most repetitive (homogeneous), followed by NW, then IM as most varied (heterogeneous).

The English and Arabic Jensen-Shannon scores look very similar and pattern similarly. For both languages when comparing for same set-size comparisons, the in-domain scores are significantly smaller than cross-domain numbers. In short, in-domain pairs are clearly identifiable by this measure in both languages.

Controlling for set size is critical: as the size of either set being compared increases, the Jensen-Shannon score

[2] The corpora sizes are all token-based, which includes punctuation in the simple tokenization condition.

decreases. This means that any comparison of Jensen-Shannon measures must take the dataset size into account. For instance, comparing a 10kx10k J-S score to a 500kx500k J-S score would lead to incorrect domain judgments.

Removing the punctuation consistently makes both in-domain and cross-domain Jensen-Shannon scores higher. This indicates that, as hypothesized, the distribution of punctuation is similar across domains, and removing the punctuation can help to tease apart content differences in domain.

There is one exception to this trend with the IM English in-domain comparisons where removing punctuation decreases the Jensen-Shannon score; this most likely speaks to the heterogeneity of this particular collection, as it consists of multiple varied instructional

English									
with punctuation					without punctuation				
		100k					100k		
		GR	NW	IM			GR	NW	IM
10k	GR	**0.268, 0.246**	0.520	0.497	10k	GR	**0.280, 0.266**	0.542	0.525
	NW	0.488	**0.253, 0.253**	0.430		NW	0.514	**0.277, 0.278**	0.471
	IM	0.485	0.450	**0.340, 0.373**		IM	0.518	0.488	**0.334, 0.358**

Table 2: English results, comparing 10k word sets to 100k word sets.

Arabic									
with punctuation					without punctuation				
		100k					100k		
		GR	NW	IM			GR	NW	IM
10k	GR	**0.242, 0.237**	0.452	0.476	10k	GR	**0.265, 0.226**	0.457	0.481
	NW	0.431	**0.229, 0.237**	0.435		NW	0.441	**0.240, 0.246**	0.436
	IM	0.468	0.448	**0.299, 0.338**		IM	0.481	0.453	**0.346, 0.356**

Table 3: Arabic results, comparing 10k word sets to 100k word sets.

manuals.[3] Furthermore, the original punctuation was more abundant in this particular corpus than in the other corpora.

Tables 2 and *3* show the results of comparing 10k word subsets to 100k word subsets for all of the domain combinations. The in-domain results appear in bold on the diagonal (there are two numbers for each of these, due to the comparison of P-10k with Q-100k, as well as P-100k with Q-10k). The in-domain Jensen-Shannon scores are, as expected, lower than the cross-domain Jensen-Shannon scores. Interestingly, the highest score here is .542 (English without punctuation, GR-10k compared to NW-100k), not even close to the Jensen-Shannon upper limit of 1.

4.2 Evaluation Phase

To implement an adaptive workflow, the objective is to route new datasets to the MT engine with whose training set they have the lowest Jensen-Shannon score. In *Tables 4* and *5*, the lowest Jensen-Shannon score for each test set appears in bold, indicating where each of these sets would be routed. The results here are largely as hypothesized, with the Jensen-Shannon measure correctly indicating that the GR, NW, and IM test sets are closest to their respective domains. Also as expected, the earlyGR test set has the lowest Jensen-Shannon score when compared to GR, and the BN test set has the lowest Jensen-Shannon score when compared to NW, with one exception. The FEMA test set has the lowest Jensen-Shannon score when compared to IM across the board, reflecting the directive, informational writing style of their bulletin-style genre across domain content.

These results are encouraging: the system would correctly route datasets from a known domain to the appropriate MT engine, if the JS measure were incorporated into a pre-MT processing component. For the domains tested, the results are consistent across different sizes, languages, and preprocessing.

5. Conclusion and Future Work

The Jensen-Shannon divergence measure is a robust method for rapidly comparing new documents by domain to known training domains and may prove effective for routing new documents to the best-available MT engine in an adaptive translation system with multiple engines.

[3] Our more recent research (publications forthcoming) has revealed that this domain is in fact quite heterogeneous. A custom MT engine trained and tested on a directed half of the set (selected using the JS scores) achieves MT accuracy (BLEU scores) similar to, and at times exceeding, that of an MT engine trained and tested on the full set.

English									
with punctuation					without punctuation				
		100k					100k		
		GR	NW	IM			GR	NW	IM
10k	GR	**0.204**	0.541	0.507	10k	GR	**0.356**	0.637	0.642
	NW	0.492	**0.238**	0.418		NW	0.518	**0.307**	0.510
	IM	0.463	0.420	**0.273**		IM	0.549	0.538	**0.414**
	earlyGR	**0.450**	0.495	0.487		earlyGR	**0.480**	0.529	0.536
	BN	0.443	**0.350**	0.368		BN	0.463	**0.365**	0.413
	FEMA	0.483	0.464	**0.413**		FEMA	0.522	0.503	**0.458**

Table 4: English results, comparing 10k word test sets with 100k word known domains.

Arabic									
with punctuation					without punctuation				
		100k					100k		
		GR	NW	IM			GR	NW	IM
10k	GR	**0.194**	0.491	0.525	10k	GR	**0.277**	0.503	0.536
	NW	0.404	**0.209**	0.418		NW	0.448	**0.236**	0.440
	IM	0.433	0.415	**0.292**		IM	0.521	0.492	**0.421**
	earlyGR	**0.347**	0.452	0.481		earlyGR	**0.348**	0.458	0.500
	BN	0.450	0.394	**0.371**		BN	0.451	**0.384**	0.392
	FEMA	0.499	0.490	**0.415**		FEMA	0.531	0.510	**0.451**

Table 5: Arabic results, comparing 10k word test sets with 100k word known domains

In this paper, we discuss removing punctuation to better pinpoint domain-ness, but expect future work on removing stop words may yield further insights. There remains the open question of how highly correlated the Jensen-Shannon scores will be to MT performance across a wider range of corpora as well. Also, this paper only examined the Jensen-Shannon divergence measure; might other divergence measures, such as the skew divergence (Lee, 2001), provide finer-tuned, more sensitive results in routing new datasets?

Given that the Jensen-Shannon scores are consistent across Arabic and English, there is an additional possibility that the Jensen-Shannon divergence measure could be used as an extrinsic evaluation of MT engines by comparing a document's source language scores with its MT output scores.

6. References

Banerjee, P.; Du, J.; Li, B.; Naskar, S.K.; Way, A.; and Genabith, J.V. (2010). Combining Multi-Domain Statistical Machine Translation Models using Automatic Classifiers. In *Proceedings of the 9th Conference of the Association for Machine Translation (AMTA 2010)*. Denver, CO.

Bird, S.; Loper, E.; and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Foster, G.; Goutte, C.; and Kuhn, R. (2010). Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*. MIT, MA, pp. 451–459.

Gildea, D. (2001). Corpus Variation and Parser Performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*. Pittsburgh, PA, pp. 167–202.

Habash, N. (2004). Large Scale Lexeme Based Arabic Morphological Generation. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN 2004)*. Fez, Morocco.

Isabelle, P.; Goutte, C.; and Simard, M. (2007). Domain Adaptation of MT Systems Through Automatic Post Editing. In *Proceedings of the Machine Translation Summit XI*. Copenhagen, Denmark, pp. 255–261.

Jaja, C.; Briesch, D; and C. Voss. (Forthcoming). Divergence Measures Tool. Internal technical report.

Kawahara, D. and Uchimoto, K. (2008). Learning reliability of parses for domain adaptation of dependency parsing. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP '08)*.

Koehn, P. and Schroeder, J. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the 2nd ACL Workshop on Statistical Machine Translation.* Prague, Czech Republic, pp. 224–227.

Kullback, S. and Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1), pp. 79–86.

Lee, L. (2001). On the Effectiveness of the Skew Divergence for Statistical Language Analysis. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics (AISTATS 2001)*. Key West, FL, pp. 65–72.

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1), pp. 145–151.

McClosky, D.; Charniak, E.; and Johnson, M. (2010) Automatic Domain Adaptation for Parsing. In

Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010). Los Angeles, CA, pp. 28–36.

Parker, R.; Graff, D.; Chen, K.; Kong, J.; and Maeda, K. (2009). *Arabic Gigaword Fourth Edition.* Linguistic Data Consortium. Philadelphia, PA.

Parker, R.; Graff, D.; Kong, J.; Chen, K.; and Maeda, K. (2011). *English Gigaword Fifth Edition*. Linguistic Data Consortium. Philadelphia, PA.

Pinto, D.; Benedi, J.M.; and Rosso, P. (2007). Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2007).* Springer-Verlag, pp. 611–622.

Plank, B. and van Noord, G. (2011). Effective Measures of Domain Similarity for Parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011).* Portland, OR.

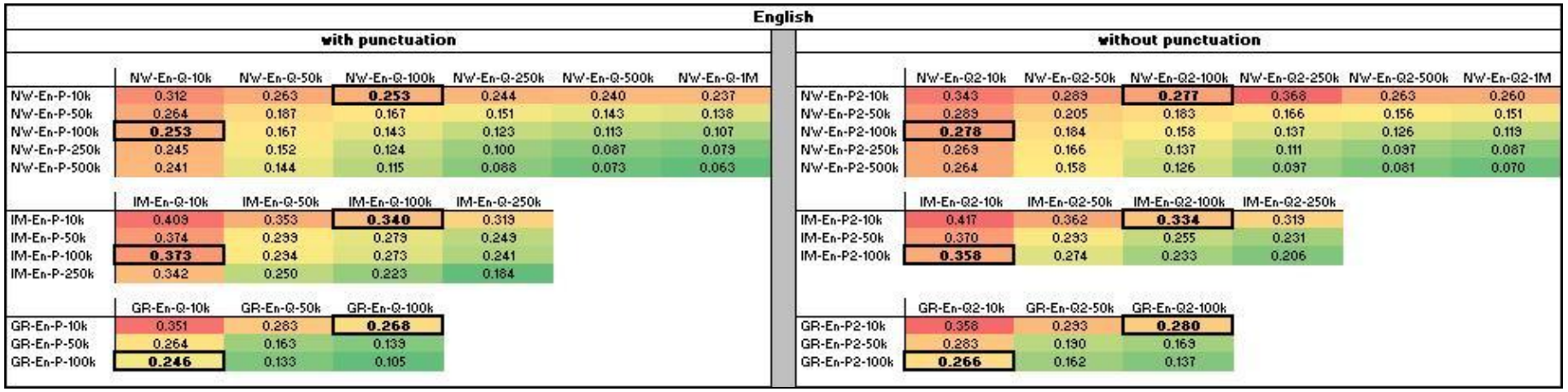
Ravi, S., Knight, K., and Soricut, R. (2008). Automatic Prediction of Parser Accuracy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008).* Honolulu, HI, pp. 887–896.

Sánchez-Martínez, F. and Forcada, M.L. (Eds.). (2011). Free/Open-Source Machine Translation. *Machine Translation (special issue),* 25(2).

Van Asch, V. and Daelemans, W. (2010). Using Domain Similarity for Performance Estimation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2010).*

Xu, J.; Deng, Y.; Gao, Y.; and Ney, H. (2007). Domain Dependent Machine Translation. In *Proceedings of the Machine Translation Summit XI.* Copenhagen, Denmark.

Zhao, B.; Eck, M.; and Vogel, S. (2004). Language Model Adaptation for Statistical Machine Translation with Structured Query Models. In *Proceedings of the International Conference on Computational Linguistics (COLING 2004).* Geneva, Switzerland, pp. 411–417.



English

with punctuation

	NW-En-Q-10k	NW-En-Q-50k	NW-En-Q-100k	NW-En-Q-250k	NW-En-Q-500k	NW-En-Q-1M
NW-En-P-10k	0.312	0.263	**0.253**	0.244	0.240	0.237
NW-En-P-50k	0.264	0.187	0.167	0.151	0.143	0.138
NW-En-P-100k	**0.253**	0.167	0.143	0.123	0.113	0.107
NW-En-P-250k	0.245	0.152	0.124	0.100	0.087	0.079
NW-En-P-500k	0.241	0.144	0.115	0.088	0.073	0.063

	IM-En-Q-10k	IM-En-Q-50k	IM-En-Q-100k	IM-En-Q-250k
IM-En-P-10k	0.409	0.353	**0.340**	0.319
IM-En-P-50k	0.374	0.299	0.279	0.249
IM-En-P-100k	**0.373**	0.294	0.273	0.241
IM-En-P-250k	0.342	0.250	0.223	0.184

	GR-En-Q-10k	GR-En-Q-50k	GR-En-Q-100k
GR-En-P-10k	0.351	0.283	**0.268**
GR-En-P-50k	0.264	0.163	0.139
GR-En-P-100k	**0.246**	0.133	0.105

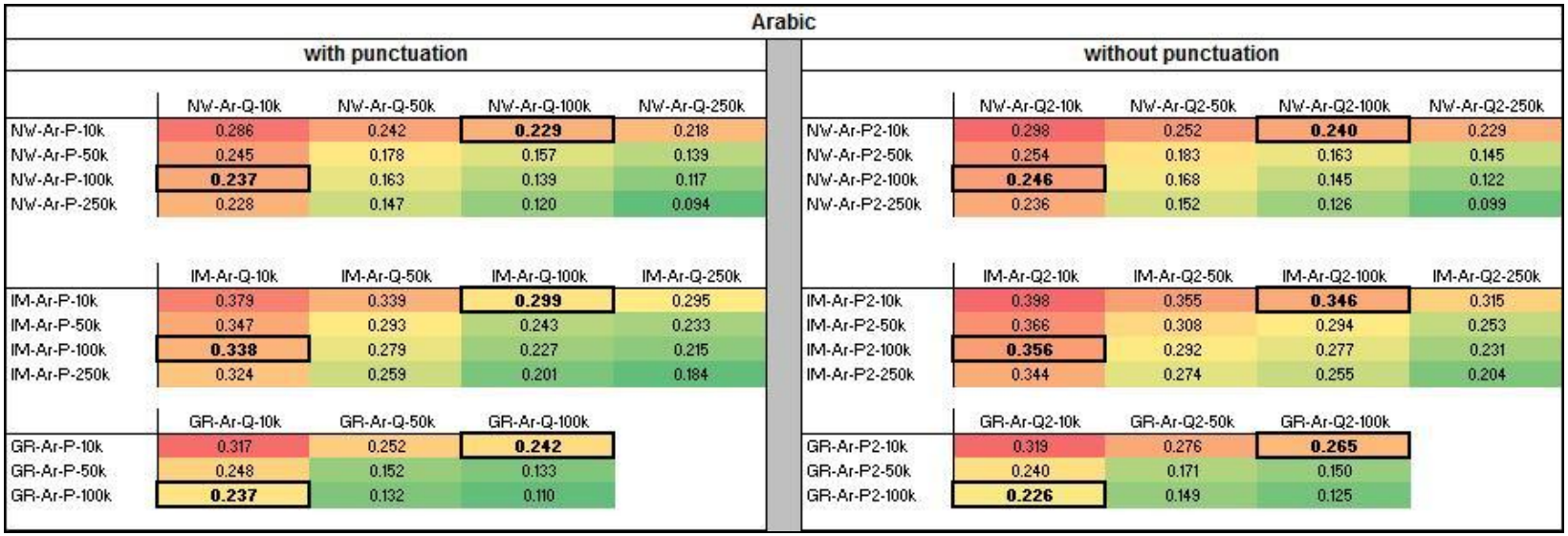
without punctuation

	NW-En-Q2-10k	NW-En-Q2-50k	NW-En-Q2-100k	NW-En-Q2-250k	NW-En-Q2-500k	NW-En-Q2-1M
NW-En-P2-10k	0.343	0.289	**0.277**	0.368	0.263	0.260
NW-En-P2-50k	0.289	0.205	0.183	0.166	0.156	0.151
NW-En-P2-100k	**0.278**	0.184	0.158	0.137	0.126	0.119
NW-En-P2-250k	0.269	0.166	0.137	0.111	0.097	0.087
NW-En-P2-500k	0.264	0.158	0.126	0.097	0.081	0.070

	IM-En-Q2-10k	IM-En-Q2-50k	IM-En-Q2-100k	IM-En-Q2-250k
IM-En-P2-10k	0.417	0.362	**0.334**	0.319
IM-En-P2-50k	0.370	0.293	0.255	0.231
IM-En-P2-100k	**0.358**	0.274	0.233	0.206

	GR-En-Q2-10k	GR-En-Q2-50k	GR-En-Q2-100k
GR-En-P2-10k	0.358	0.293	**0.280**
GR-En-P2-50k	0.283	0.190	0.163
GR-En-P2-100k	**0.266**	0.162	0.137

Figure 2: English in-domain Jensen-Shannon scores. The color scale from green to red indicates the lowest (least divergent) to highest (most divergent) scores. The bolded cells are the scores used for *Table 2*.



Arabic

with punctuation

	NW-Ar-Q-10k	NW-Ar-Q-50k	NW-Ar-Q-100k	NW-Ar-Q-250k
NW-Ar-P-10k	0.286	0.242	**0.229**	0.218
NW-Ar-P-50k	0.245	0.178	0.157	0.139
NW-Ar-P-100k	**0.237**	0.163	0.139	0.117
NW-Ar-P-250k	0.228	0.147	0.120	0.094

	IM-Ar-Q-10k	IM-Ar-Q-50k	IM-Ar-Q-100k	IM-Ar-Q-250k
IM-Ar-P-10k	0.379	0.339	**0.299**	0.295
IM-Ar-P-50k	0.347	0.293	0.243	0.233
IM-Ar-P-100k	**0.338**	0.279	0.227	0.215
IM-Ar-P-250k	0.324	0.259	0.201	0.184

	GR-Ar-Q-10k	GR-Ar-Q-50k	GR-Ar-Q-100k
GR-Ar-P-10k	0.317	0.252	**0.242**
GR-Ar-P-50k	0.248	0.152	0.133
GR-Ar-P-100k	**0.237**	0.132	0.110

without punctuation

	NW-Ar-Q2-10k	NW-Ar-Q2-50k	NW-Ar-Q2-100k	NW-Ar-Q2-250k
NW-Ar-P2-10k	0.298	0.252	**0.240**	0.229
NW-Ar-P2-50k	0.254	0.183	0.163	0.145
NW-Ar-P2-100k	**0.246**	0.168	0.145	0.122
NW-Ar-P2-250k	0.236	0.152	0.126	0.099

	IM-Ar-Q2-10k	IM-Ar-Q2-50k	IM-Ar-Q2-100k	IM-Ar-Q2-250k
IM-Ar-P2-10k	0.398	0.355	**0.346**	0.315
IM-Ar-P2-50k	0.366	0.308	0.294	0.253
IM-Ar-P2-100k	**0.356**	0.292	0.277	0.231
IM-Ar-P2-250k	0.344	0.274	0.255	0.204

	GR-Ar-Q2-10k	GR-Ar-Q2-50k	GR-Ar-Q2-100k
GR-Ar-P2-10k	0.319	0.276	**0.265**
GR-Ar-P2-50k	0.240	0.171	0.150
GR-Ar-P2-100k	**0.226**	0.149	0.125

Figure 3: Arabic in-domain Jensen-Shannon scores. The color scale from green to red indicates the lowest (least divergent) to highest (most divergent) scores. The bolded cells are the scores used for *Table 3*

English

with punctuation

	NW-En-10k	NW-En-50k	NW-En-100k	NW-En-250k	NW-En-500k	NW-En-1M
IM-En-10k	0.481	0.456	**0.450**	0.446	0.444	0.443
IM-En-50k	0.442	0.402	0.392	0.385	0.381	0.378
IM-En-100k	**0.430**	0.384	0.373	0.364	0.359	0.356
IM-En-250k	0.414	0.362	0.348	0.337	0.331	0.327
IM-En-500k	0.402	0.345	0.330	0.317	0.310	0.305

	IM-En-10k	IM-En-50k	IM-En-100k	IM-En-250k	IM-En-500k
GR-En-10k	0.538	0.507	**0.497**	0.485	0.472
GR-En-50k	0.495	0.451	0.436	0.418	0.402
GR-En-100k	**0.485**	0.437	0.421	0.401	0.383

	GR-En-10k	GR-En-50k	GR-En-100k
NW-En-10k	0.546	0.499	**0.488**
NW-En-50k	0.525	0.465	0.449
NW-En-100k	**0.520**	0.456	0.440
NW-En-250k	0.516	0.450	0.432
NW-En-500k	0.514	0.446	0.428
NW-En-1M	0.513	0.445	0.426

without punctuation

	NW-En-2-10k	NW-En-2-50k	NW-En-2-100k	NW-En-2-250k	NW-En-2-500k	NW-En-2-1M
IM-En-2-10k	0.521	0.494	**0.488**	0.484	0.482	0.480
IM-En-2-50k	0.479	0.437	0.426	0.418	0.414	0.411
IM-En-2-100k	**0.471**	0.423	0.412	0.402	0.397	0.394
IM-En-2-250k	0.466	0.412	0.398	0.386	0.380	0.375

	IM-En-2-10k	IM-En-2-50k	IM-En-2-100k	IM-En-2-250k
GR-En-2-10k	0.567	0.533	**0.525**	0.526
GR-En-2-50k	0.528	0.480	0.468	0.465
GR-En-2-100k	**0.518**	0.465	0.452	0.447

	GR-En-2-10k	GR-En-2-50k	GR-En-2-100k
NW-En-2-10k	0.571	0.527	**0.514**
NW-En-2-50k	0.547	0.489	0.472
NW-En-2-100k	**0.542**	0.480	0.461
NW-En-2-250k	0.537	0.473	0.453
NW-En-2-500k	0.535	0.469	0.448
NW-En-2-1M	0.534	0.468	0.446

Figure 4: English cross-domain Jensen-Shannon scores. The color scale from green to red indicates the lowest (least divergent) to highest (most divergent) scores. The bolded cells are the scores used for *Table 2*.

Arabic

with punctuation

	NW-Ar-10k	NW-Ar-50k	NW-Ar-100k	NW-Ar-250k	NW-Ar-500k
IM-Ar-10k	0.475	0.453	**0.448**	0.443	0.441
IM-Ar-50k	0.445	0.412	0.403	0.395	0.391
IM-Ar-100k	**0.435**	0.398	0.387	0.377	0.372
IM-Ar-250k	0.418	0.374	0.361	0.349	0.342
IM-Ar-500k	0.407	0.360	0.345	0.332	0.324

	IM-Ar-10k	IM-Ar-50k	IM-Ar-100k	IM-Ar-250k	IM-Ar-500k
GR-Ar-10k	0.510	0.485	**0.476**	0.460	0.450
GR-Ar-50k	0.475	0.437	0.424	0.402	0.388
GR-Ar-100k	**0.468**	0.427	0.412	0.387	0.372

	GR-Ar-10k	GR-Ar-50k	GR-Ar-100k
NW-Ar-10k	0.483	0.439	**0.431**
NW-Ar-50k	0.459	0.402	0.389
NW-Ar-100k	**0.452**	0.390	0.376
NW-Ar-250k	0.446	0.380	0.364
NW-Ar-500k	0.443	0.375	0.358

without punctuation

	NW-Ar-2-10k	NW-Ar-2-50k	NW-Ar-2-100k	NW-Ar-2-250k	NW-Ar-2-500k
IM-Ar-2-10k	0.481	0.458	**0.453**	0.449	0.446
IM-Ar-2-50k	0.450	0.415	0.405	0.398	0.393
IM-Ar-2-100k	**0.436**	0.396	0.385	0.375	0.369
IM-Ar-2-250k	0.424	0.378	0.365	0.353	0.345
IM-Ar-2-500k	0.412	0.362	0.347	0.333	0.325

	IM-Ar-2-10k	IM-Ar-2-50k	IM-Ar-2-100k	IM-Ar-2-250k	IM-Ar-2-500k
GR-Ar-2-10k	0.519	0.493	**0.481**	0.470	0.459
GR-Ar-2-50k	0.490	0.452	0.435	0.419	0.404
GR-Ar-2-100k	**0.481**	0.439	0.420	0.401	0.384

	GR-Ar-2-10k	GR-Ar-2-50k	GR-Ar-2-100k
NW-Ar-2-10k	0.489	0.452	**0.441**
NW-Ar-2-50k	0.463	0.414	0.397
NW-Ar-2-100k	**0.457**	0.403	0.384
NW-Ar-2-250k	0.451	0.394	0.373
NW-Ar-2-500k	0.448	0.388	0.366

Figure 5: Arabic cross-domain Jensen-Shannon scores. The color scale from green to red indicates the lowest (least divergent) to highest (most divergent) scores. The bolded cells are the scores used for *Table 3*.