



## NRC Publications Archive Archives des publications du CNRC

### Filtering and routing multilingual documents for translation

Carpuat, Marine; Goutte, Cyril; Isabelle, Pierre

For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

#### **Publisher's version / Version de l'éditeur:**

<http://dx.doi.org/10.1109/CISDA.2012.6291536>

*2012 IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA 2012), pp. 1-7, 2012-07-13*

#### **NRC Publications Record / Notice d'Archives des publications de CNRC:**

<http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?lang=en>

<http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?lang=fr>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

[http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/jsp/nparc\\_cp.jsp?lang=en](http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/jsp/nparc_cp.jsp?lang=en)

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

[http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/jsp/nparc\\_cp.jsp?lang=fr](http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/jsp/nparc_cp.jsp?lang=fr)

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Contact us / Contactez nous: [nparc.cisti@nrc-cnrc.gc.ca](mailto:nparc.cisti@nrc-cnrc.gc.ca).



# Filtering and Routing Multilingual Documents for Translation

Marine Carpuat, Cyril Goutte and Pierre Isabelle

Interactive Language Technology  
National Research Council Canada  
Gatineau, QC J8X3X7 Canada  
Email: Firstname.Lastname@nrc.ca

**Abstract**—Translation is a key capability to access relevant information expressed in various languages on social media. Unfortunately, systematically translating all content far exceeds the capacity of most organizations. Computer-aided translation (CAT) tools can significantly increase the productivity of translators, but can not ultimately cope with the overwhelming amount of content to translate. In this contribution, we describe and experiment with an approach where we use the structure in a corpus to adequately route the content to the proper workflow, including translators, CAT tools or purely automatic approaches. We show that linguistically motivated structure such as document genre can help decide on the proper translation workflow. However, automatically discovered structure has an effect that is at least as important and allows us to define groups of documents that may be translated automatically with reasonable output quality. This suggests that computational intelligence models that can efficiently organize document collection will provide increased capability to access textual content from various target languages.

## I. INTRODUCTION

Organizations interested in monitoring social media for commercial or security purposes must handle growing amounts of content, but also an increasing number of languages. Accessing multilingual information is a key capability for analyzing this data efficiently. However, traditional translation performed by professional human translators suffers from two main drawbacks in the context of social media data analysis. First it is a slow process that would bottleneck automatic analysis in the large text collections generated by social media users. Second, and more importantly, systematic translation of multilingual content far exceeds the capacity of most organizations. Although many tools are available to ease the translators job, such as translation memories, terminology banks or bilingual concordancers, the gains in productivity cannot match the growth in content that needs to be translated.

Machine translation promises to bring productivity gains, but suffers from a number of problems. Although its output is often good enough for gisting, it is still far from reaching the same quality as human translation. More problematic is the fact that state-of-the-art, statistical machine translation (SMT) is highly sensitive to mismatches between the text to translate and the training material used to estimate the translation model. As a consequence, text that is unlike the training data may produce poor translations. Typical experimental settings considered in SMT research (e.g. [1], [2]) use well defined

benchmark data in a well controlled setting. In contrast, social media data is highly dynamic in nature, with many different sources spanning a range of topics, genres and quality, such that appropriate representative training data is hard to find.

We address the following questions. First, how does translation quality of computer-aided translation (CAT) tools vary across text genres? Heterogeneous text collections can be organized along many dimensions. We propose to start with genre, since the genre of a document attempts to represent its medium and communicative goal. Genre is therefore particularly relevant in the context of social media, where documents vary widely depending on their source and purpose, and are quite different from the texts that have been historically studied in machine translation (e.g., Canadian or European parliament proceedings [3]). We will show that there are systematic differences in CAT performance across genres. This suggests that genre information (either given or automatically attributed [4]) is useful to route documents to be translated to the appropriate CAT tools.

The second question we address is whether we can automatically discover relevant structure from heterogeneous data in order to translate a wide variety of documents more efficiently and accurately. We show that indeed, we discover clusters of textual documents on which CAT tools have very different and consistent behaviour. This suggests a natural way to route documents to the appropriate combination of fully automatic translation, semi-automatic translation involving, e.g. post-edition, or fully human translation when CAT tools are ineffective.

As there is currently no publicly available large corpus of translated social media data, we illustrate our approach and results on a multilingual document collection from the United Nations. This data set offers several realistic and desirable features: it is large, it covers several languages, and it contains text of very different genres. Contrary to most data used in security and defence applications, it is publicly available and can be used to demonstrate the approach.

In the following two sections, we describe the methods used in our approach. Section IV describes the data that we use in our experiments, and sections VI and VII shows how corpus structure may be leveraged to improve the translation workflow, using both existing manual distinctions, as well as automatically discovered corpus structure.

## II. METHODS: COMPUTER-AIDED TRANSLATION

This section gives a brief overview of the two main computer-aided translation methods that are widely used today: translation memories and statistical machine translation. Since our contribution is focused on evaluation of translation quality and its impact on routing documents for translation, we are not interested in proposing new methods here. On the contrary, we aim to use well-established tools in order to build a pipeline that is representative of current practice.

### A. Translation Memory

The first type of CAT tool we consider in this study are Translation Memories (TM). While they have not been studied much in the field of machine translation, TM are widely used by professional translators. TM improve translators' productivity and consistency by providing examples of similar sentences that have already been translated.

We use an NRC-internal implementation [5] that simulates commercial systems while giving us more control over parameters. Given a corpus of existing translated sentences, and a new source sentence that we call the *query*, we compute the *distance* between the query and every source sentence in the corpus. The translation memory outputs the target side of the sentence pair with the minimum distance between its source side and the query. We report results using  $1 - \text{smoothedBLEU}$  as the distance metric, where *smoothedBLEU* is a smoothed version of BLEU score [6], the standard metric of translation quality based on  $n$ -gram precision between system output and reference translations [7]. We experimented with other distance metrics and found that *smoothedBLEU* worked best when using BLEU as the evaluation metric.

### B. Statistical Machine Translation

Our second CAT tool is Statistical Machine Translation (SMT). In contrast with TM, SMT is a very active topic of research and a wide variety of models and systems exist. We use the NRC's PORTAGE system [8], a state-of-the-art phrase-based SMT system which placed 3rd out of 19 participants in the constrained Chinese-to-English task at the 2009 NIST OpenMT evaluation, and performed very well in the latest 2012 evaluation for both Chinese-English and Arabic-English. In this paper, the SMT decoder scores translation candidates for each sentence using a loglinear mixture of the standard set of features (conditional translation probabilities and lexical weights for the translation model, lexicalized distortion and target language model). Features and their weights are estimated from bilingual parallel text. We refer the reader to [9] for a good overview of phrase-based SMT modeling.

## III. METHODS: CORPUS ORGANIZATION

We need tools to automatically learn how to organize our large heterogeneous collection of documents. We use a two step clustering approach for this purpose: (1) a Latent Dirichlet Allocation (LDA) model is used to discover latent topics in the document collection, and (2) documents are then

clustered using their topic distribution as a low-dimensional representation.

We use the MALLET<sup>1</sup> implementation of LDA [10] to learn topics in a document collection. LDA is a generative model of documents, where a topic  $t$  is a distribution over words  $\phi_t$ , and each document  $d$  is a mixture of  $T$  latent topics, given by probabilities  $\theta_d$ . Dirichlet priors are placed over both  $\phi$  and  $\theta$ . In a document  $d$ , the word tokens  $w^d = \{w_n^d\}_{n=1}^{N_d}$  are associated with topic assignments  $z^d = \{z_n^d\}_{n=1}^{N_d}$ , drawn from the document-specific topic distribution. Given a corpus of observed words  $w$ , the posterior distribution over  $z$  can be estimated using Gibbs sampling, and used to compute the word-topic distributions and the topic-document distribution. In practice, we run the sampler for 1000 iterations. The concentration parameter  $\beta$  for the topic-word concentration is initialized to 0.01, while the document-specific topic distributions are initialized with a concentration parameter  $\alpha = 0.01/T$  and a uniform base measure [11]. Hyperparameters are reestimated every 10 sampling iterations.

Each document in the corpus can be represented as a vector  $d \in [0, 1]^T$ , where  $T$  is the number of topics and  $d$  is the document-specific topic distribution from the LDA model. The resulting topic vectors are clustered using the  $k$ -means algorithm, as implemented in the WEKA toolkit[12]. The main parameters are therefore the number of topics  $T$  and the number of clusters  $k$ . In this paper, we use  $k = 8$ , so that we have the same number of genres and clusters, and we experimented with a range of topic numbers. We will report results with  $T = 80$ , and will give an overview of topics discovered in our data in Section VII.

The resulting two-step approach of LDA followed by  $k$ -means is representative of current approaches to document categorization [13, for a good overview]. It remains to be seen whether purely Computational Intelligence (CI) techniques such as fuzzy or neural systems can improve models for this task. In this contribution, we focus on evaluating the impact of standard approaches, and leave the investigation of more advanced CI models to future work.

## IV. DATA: A HETEROGENEOUS MULTILINGUAL CORPUS

We work with a bilingual corpus of Chinese-English documents extracted from United Nations (UN) data and made available by the Linguistic Data Consortium.<sup>2</sup> The corpus comprises two parts:

- 1) A genre annotated subset of 972 Chinese-English documents, all published in 2000, with a genre label for each document. We refer to this subset as UN-2000.
- 2) An unannotated, parallel, sentence-aligned corpus of 33,174 Chinese-English documents published between 1993 and 1999. We refer to this subset as UN-9399.

Note that all documents come from the same general source (the United Nations), but the dates of the two subsets do

<sup>1</sup><http://mallet.cs.umass.edu/>

<sup>2</sup><http://www.ldc.upenn.edu/>

TABLE I  
GENRE CATEGORIES IN THE DATASET

#	Title	Description
A	Biography/Résumé	Account of persons life; résumé/CV
B	Lecture/Briefing	Formal statement presented to committee; one side of debate, speech
C	Letter	Communication introduced by word "Letter" with addressee; line 1 stating "I have the honor to inform you that"; ending with signature
D	Meeting Documentation	Agenda, addendum/addenda, Minutes, Summary of Meeting; references to Draft Resolution(s), statements, reports, and comments
E	Draft Resolution	Detailed quotation of resolution or draft; series of paragraphs beginning with gerunds, then requests/decides (3rd s./present verbs.); allusion to session agenda
F	Message	Short or long communication, including the note verbale, written note by official (instruction or brief modification to agenda item/addendum)
G	Organizational Charter	Articles of organization, org. chart/lists, provisions for procedures and responsibilities; guidance and rules for group/members; job description(s)/vacancy announcement(s)
H	Study/Report	Detailed official statements, conclusions and recommendations; progress report and summary of issues on given topic

not overlap. This ensures that no document from the genre-annotated corpus is present in the large sentence-aligned corpus. This also ensures that the genre annotated documents come from the same source, but may cover new topics that appeared in 2000 but were not necessarily represented in previous years, a situation that is common in typical applications.

The genre annotation was originally collected by collaborators for an independent sentence alignment project. The list of genres is given in table I with a brief description of each. These categories are quite different from genres used in the traditional genre categorization literature (see e.g. [4] and references therein) and most are specific to the UN data. Some are defined by the communication medium, or by rules based on text patterns (e.g., Genre C, Letter). While some genres cover a wide variety of documents (G, org. charter), others make very fine grained distinctions: for instance, agenda and addenda should be in "Meeting Documentation" (D), while "Draft Resolution" (E) contains allusions to session agenda, and "Message" (F) contains written notes by officials which can be brief modifications to agenda items or addendum.

Each document was annotated independently by two human judges, and a third judge was asked to adjudicate documents with a final reference annotation. It is important to note that the disagreement between judges was fairly high, with an agreement ratio between 0.44 and 0.76 depending on the pair of judges, and a *kappa* between 0.31 and 0.70. This suggests that the genre annotation task is non trivial, and proved difficult for human judges, even with precise guidelines and definitions (cf. Table I). We therefore expect that genre cues are not simple to identify in text.

This annotation process yielded a total of 972 documents, each annotated with one of eight genres, which we refer to as UN-2000. Note that the distribution of documents per genre is

TABLE II  
TEXT STATISTICS PER GENRE FOR UN-2000 CORPUS: NUMBER OF DOCUMENTS, NUMBER OF SENTENCES PER DOCUMENT; FOR EACH LANGUAGE: AVERAGE SENTENCE LENGTH ( $\bar{\ell}$ ), NUMBER OF SINGLETONS PER DOCUMENT (SNGL), AND TOKEN TO TYPE RATIO (TTR).

#	nb.		Chinese			English		
	docs	sent /doc	$\bar{\ell}$	sngl	ttr	$\bar{\ell}$	sngl	ttr
A	5	11.2	13.5	46.2	2.05	15.0	47.2	2.41
B	7	95.7	24.1	143.0	6.94	26.9	199.6	6.41
C	190	10.1	22.7	10.7	9.68	25.6	14.1	9.09
D	264	67.8	23.4	16.5	37.66	26.0	22.7	31.50
E	165	29.3	30.3	12.9	25.14	33.3	18.0	21.96
F	95	13.5	19.4	16.2	7.39	21.4	19.6	7.39
G	26	111.7	19.7	54.2	15.18	21.3	75.9	13.24
H	215	111.1	25.3	28.6	40.85	27.3	44.6	30.82

very unbalanced (Table II). The biography (A) and lecture (B) genres have fewer than 10 instances each, while the meeting documentation (D) genre alone represents more than 25% of all annotated documents. The length of documents vary with genre, but the number of sentences per document reported in Table II do not reflect the original document length, since sentences were discarded for the sentence alignment project.

Table II gives basic statistics per genre in both languages: average sentence length, number of singletons (words occurring only once in the corpus) per document, and token-to-type ratio (or average word frequency). The biography/résumé genre (A) for instance has shorter sentences than average, while meeting documentation (D) has much longer sentences. Genres based on more formal or official documents tend to reuse more vocabulary with high token to type ratio (D, E, H), while informal genres exhibit more variation in vocabulary (B, C, F). The biography/résumé genre also has a very low token to type ratio, which can be explained by its small size and by the frequent occurrence of named-entities. As can be expected, the same trends are observed in Chinese and English.

It is hard to predict a priori how computer-aided translation tools are affected by these properties: it depends on the translation technology used, as well as on the properties of the existing translations used as a translation memory or training corpus. A statistical machine translation system might make fewer reordering errors in shorter sentences. However, the short sentences of genre A might still be hard to translate correctly if they contain many rare words that have not been seen in the training data. In a translation memory, short sentences might yield incorrect translations due to lack of context, while long sentences with many repetitions (such as in genre E) might be easier to translate if they are well covered by the examples in the memory.

In the following experiments, the large, non-annotated corpus is used to feed the translation memory and to train the statistical machine translation engine. The smaller, genre-annotated UN2000 corpus is used to test the impact of the corpus structure on the document translation quality. In particular, we check the impact of the genre on the translation quality.

## V. EXPERIMENT SET-UP

We consider the two types of CAT tools described in Section II: translation memories (TM) and statistical machine translation (SMT). The TM and SMT systems are built using the exact same training data. We use the 33k documents from the UN-9399 corpus. This results in a training set of more than 3 Million sentence pairs, which is on par with the large-scale conditions considered in machine translation benchmarks.

The UN-2000 section of the corpus is used to evaluate the performance of the CAT tools. In order to speed up experimentation, we do not use the entire genre-annotated corpus as a test set, but only a subset of 51 randomly selected documents, or a total of 2942 sentences which is roughly the standard size for SMT test sets.

We will evaluate translation quality on different partitions of the corpus: (1) genres, according the manual annotation described in Section IV, and (2) document clusters learned automatically as described in Section III. Note that translations are always produced using a single CAT tool, i.e., a single translation memory or a single machine translation system trained on the entire training set. The resulting CAT tool is then applied to all test documents. Genre and cluster information does not inform the translation process, and is only used at evaluation time.

We use BLEU score [7] to automatically compare TM and SMT output with reference translations. Although some recent metrics have been shown to correlate slightly better with human judgments of translation quality, BLEU remains the de-facto standard metric in statistical MT, both for system optimization and for evaluation [14], [1]. BLEU is defined as the geometric mean of  $n$ -gram precision scores for  $n = 1..4$ , augmented with a length penalty, so that higher BLEU scores represent better translation quality. We did experiment with other metrics such as Word Error Rate [15], but do not include them because they yield the same trends as BLEU.

In addition to the overall BLEU score obtained for each partition of the UN-2000 corpus, we will compute BLEU scores for each document in the corpus. When comparing the translation quality of different systems on the same test set, corpus-level BLEU score is known to correlate better with human judgments of translation quality than document-level scores. However, document-level BLEU scores are still interesting in our setting since we are not interested in comparing different SMT systems on a fixed data set, but in comparing translation quality of a fixed SMT system across different data sets. Following recent work on translation quality estimation [16], we study the distribution of documents over BLEU quartiles rather than on the absolute BLEU values alone. We rank all documents in the UN-2000 test set according to their BLEU score, and report for each genre the percentage of documents that are ranked in each quartile.

## VI. IMPACT OF GENRE STRUCTURE ON CAT

In this section, we evaluate standard CAT tools based on genre distinctions.

TABLE III  
BLEU SCORES FOR COMPUTER-AIDED TRANSLATION, BY GENRE.

Genre	# sent	TM BLEU	SMT BLEU
A	3	54.25	66.99
B	17	19.92	42.44
C	50	13.75	24.56
D	292	19.39	35.94
E	892	50.66	52.30
F	21	58.72	58.37
G	52	39.91	53.60
H	1635	23.94	32.28
All:	2942	27.29	36.97

TABLE IV  
DOCUMENT-LEVEL BLEU SCORES BY GENRE. BLEU QUANTILES ARE DEFINED BY RANKING DOCUMENTS ACROSS ALL GENRES, AND ARE ORDERED FROM BEST (1ST) TO WORST (4TH)

Genre	BLEU by doc			docs in each BLEU quartile (%)			
	Mean	Min	Max	1st	2nd	3rd	4th
A	66.99	66.99	66.99	1	0	0	0
B	42.44	42.44	42.44	0	1	0	0
C	32.32	17.06	51.30	0	0.25	0.25	0.5
D	44.11	22.68	65.71	0.307	0.153	0.384	0.153
E	44.53	26.12	57.32	0.285	0.285	0.428	0
F	54.62	42.44	61.80	0.666	0.333	0	0
G	55.24	47.41	63.08	0.5	0.5	0	0
H	44.37	22.76	70.45	0.230	0.307	0.307	0.153

### A. Genres capture some useful distinctions for CAT

Table III summarizes the BLEU scores obtained on each genre subset of the UN-2000 test set. It shows that CAT tools do very well on some genres such as E and F. This was perhaps expected for genre E, as we have seen in Section IV that it is one of the most formal and repetitive genres. Interestingly, genre F which is more informal and less repetitive also gets a high BLEU score even though it might be considered hard to translate a priori. Genre A also shows a high BLEU score, but as this is the smallest genre, these scores are probably not reliable. The worst BLEU scores with both CAT tools are obtained for genre C.

Statistics for BLEU scores per documents are reported in Table IV. While the mean document-BLEU do not yield the exact same ranking as the genre-based BLEU, the top and worst genres remain the same. The distribution of documents in each BLEU quartile suggests that translations of documents in genres F and G are more reliable than those in genres E or C, since all documents in genres F and G are ranked within the best two BLEU quartiles. In contrast, half of the documents in genre C are in the worst BLEU quartile, confirming that its translation quality is not as good as those of the other genres.

### B. Genres capture non-trivial distinctions

Could we have guessed which genres are easier to translate simply by comparing test documents with the training data? Unfortunately genre annotation is not available for the training corpus, so we cannot directly compare the genre distributions in the training and test data. However, we can compare the Chinese documents in each genre with the entire training corpus using Out-Of-Vocabulary (OOV) rate and source language model perplexity per word (PPL). Given a sequence of

TABLE V  
LANGUAGE MODEL EVALUATION PER GENRE: OUT OF VOCABULARY  
WORD RATE (OOV) AND PERPLEXITY (PPL) FOR CHINESE AND ENGLISH  
4-GRAM LANGUAGE MODELS TRAINED ON THE UN93-99 CORPUS

Genre	Chinese		English	
	OOV	PPL	OOV	PPL
A	10.61	336.76	4.28	191.51
B	0.17	107.35	0.21	80.64
C	1.57	126.58	1.27	93.81
D	0.79	106.72	0.54	74.63
E	0.87	77.28	0.50	57.95
F	2.04	163.70	1.25	110.61
G	0.72	186.93	1.07	124.68
H	0.71	156.16	0.74	109.65

words  $S = (w_1, \dots, w_s)$  (e.g, a sentence, document, or set of documents), they are defined as follows:

- the out-of-vocabulary rate, *OOV*, represents the ratio of test tokens that are unseen in the UN93-99 corpus:

$$OOV = \frac{\sum_{i=1}^s (1 - \mathbb{1}_{UN93-99}(w_i))}{s}$$

- the perplexity per word, *PPL*, represents the amount by which the  $n$ -gram language model  $LM^3$  reduces uncertainty about the next word:

$$\log(PPL) = -\frac{1}{s} \sum_{i=1}^s \log P_{LM}(w_i | w_{i-n+1}^{i-1})$$

While the genre with the lowest PPL is the one that gets the higher BLEU score, OOV and PPL alone are not sufficient to predict translation performance. In particular, genre F yields one of the highest BLEU scores despite the high perplexity of the Chinese language model on this genre. This suggests that predicting translation quality is not straightforward and that genre distinctions do capture useful distinctions that cannot be predicted by the similarity of test and training documents alone.

### C. SMT and TM behave differently across genres

Comparing results across CAT tools, Table III shows that there is a bigger variability in behaviour with TM than with SMT: the performance on the three worst genres is much lower for the TM. In fact, it is on these genres that the SMT outperforms the TM most, at least according to the BLEU metric. Overall, it is clear that the SMT consistently does at least as well as the TM, and sometimes much better.

One key to explaining this difference lies in the distinct ways in which the two tools function. A TM performs well when its *coverage* is good: there are many segments in the text to translate that match or are very similar to the memorized training data. On genres where the coverage is lower, the best the TM can do is return the target side of poorly-matching segments. On the contrary, the SMT system is able to generate translations that have never been seen in the training corpus. SMT will obviously perform better when the test set is close to its training data, but when it is not, SMT at least attempts to

produce a prediction that contains translations of the source segment, often at the expense of grammaticality, instead of just outputting a grammatical, but poorly-related segment.

Another factor is that SMT and TM are affected differently by alignment errors in the training data. Since our training corpus was obtained by automatically aligning sentence pairs (which is standard procedure in SMT), there is a significant number of sentence alignment errors. Due to the way the SMT system is estimated, these alignment errors tend to be “averaged out” by the statistical estimation process. On the other hand, the TM will just return a misaligned target segment as long as the source segment matches. TM is therefore less robust to alignment errors than SMT.

### D. Impact on translation workflow

This evaluation of translation quality suggests that organizing documents by genre is useful in order to decide how to route documents for translation. Our results suggest that different strategies might be needed depending on genres. For instance, we can have a two-stage strategy where CAT tools alone are used to translate the best performing genres (E and F). SMT can be used either alone or in combination with TM. For genres with lower BLEU scores and more documents in the lower BLEU quartiles, SMT will probably require human post-edition in order to be usable.

## VII. IMPACT OF LEARNED CORPUS STRUCTURE ON CAT

After evaluating translation quality by genre, we now perform a similar evaluation after automatically partitioning the training and test documents in an unsupervised fashion. Clusters are obtained automatically using the combination of unsupervised topic modeling and clustering described in Section III. We set the number of clusters  $K = 8$ , in order to have the same number of clusters and genres, and  $T = 80$  after experimenting with various numbers of topics.

While it is unclear how to evaluate quality of topics intrinsically, manual inspection reveals that even with a relatively high number of topics the distinctions captured seem reasonably well motivated. We ranked words in each LDA topic according to the absolute value of their contribution to the Kullback Leibler divergence between each topic distribution considered and the word distribution in the entire UN9399 corpus. In order to give the reader a sense of the topics captured, we report here the top 10 words for the first few topics learned:

- topic 1:** general, committee, resolution, assembly, united, council, secretary, draft, report, session
- topic 2:** government, security, council, republic, special, military, refugees, united, humanitarian, situation
- topic 3:** staff, period, general, space, cost, united, amount, costs, cent, mission
- topic 4:** development, countries, women, economic, social, developing, international, national, trade, world
- topic 5:** nations, united, programme, activities, general, development, information, system, organizations, support
- topic 6:** rights, human, international, united, convention, states, nations, special, commission, children

<sup>3</sup>In our experiments we use  $n = 5$ , i.e. 5-gram language model.

TABLE VI  
BLEU SCORES FOR COMPUTER-AIDED TRANSLATION, BY CLUSTER.

Cluster	# sent	TM BLEU	SMT BLEU
1	221	54.97	54.61
2	195	12.51	33.18
3	35	60.68	54.25
4	7	0.00	0.00
5	101	62.98	61.82
6	181	14.30	34.42
7	427	16.82	36.48
8	1775	24.09	31.65
All:	2942	27.29	36.97

TABLE VII  
DOCUMENT-LEVEL BLEU SCORES BY CLUSTER. BLEU QUANTILES ARE DEFINED BY RANKING DOCUMENTS ACROSS ALL GENRES, AND ARE ORDERED FROM BEST (Q1) TO WORST (Q4). CLUSTER 4 IS TOO SMALL TO COMPUTE STATISTICS.

Cluster	BLEU by doc			% docs per BLEU quartile			
	Mean	Min	Max	q1	q2	q3	q4
1	54.61	54.61	54.61	1	0	0	0
2	33.30	24.39	42.44	0	0.2	0.6	0.2
3	48.39	25.38	61.80	0.2	0.6	0	0.2
4	-	-	-	-	-	-	-
5	60.15	53.75	66.99	0.714	0.285	0	0
6	34.07	33.25	34.89	0	0	1	0
7	41.88	30.96	54.68	0.166	0.333	0.5	0
8	42.85	17.06	70.45	0.277	0.222	0.277	0.222

- **topic 7:** article, state, law, international, court, convention, states, legal, paragraph, commission

#### A. Learned clusters capture useful distinctions for CAT

The BLEU scores per cluster summarized in Table VI are overall fairly similar to what we observed on the genre-based partitioning in Section VII.

CAT tools perform very well on some clusters (e.g., clusters 1, 3 and 5), less on others (2,4,6,8). As a result, clusters do identify useful distinctions for the purpose of routing documents. Just like with genres, the Translation Memory performs poorly on some clusters, where the SMT performance is also lower, but much better than that of the TM. Both TM and SMT do poorly on cluster 4, which happens to be very small and mostly composed of “documents” which contain list of numbers and other non-words which should probably not be “translated”, in the usual sense of the word. If anything, it seems that clusters reinforce what we saw on genres: on the best performing clusters, the TM does even better than the SMT, while on the weaker clusters SMT performance is much less impacted.

As for genres, we study the document-level BLEU scores for each cluster in Table VII. These scores also highlight the good translation quality of cluster 5, which has by far the highest mean score, and more than 70% of documents ranked within the top BLEU quartile.

#### B. Clusters capture different distinctions than genres

Automatically learned clusters are quite different from genres. The cluster distribution is more balanced than the genre distribution overall. In addition, the distribution of test

TABLE VIII  
TOPIC-BASED CLUSTERS AND GENRES RESULT IN VERY DIFFERENT PARTITIONS OF THE UN2000 DOCUMENTS, AS QUANTIFIED BY PURITY, RAND INDEX (RI), PRECISION, RECALL AND  $F_1$  SCORE

purity	RI	p	r	$F_1$
71.75	74.06	34.31	34.59	34.45

documents across genres for each cluster shows that there is not a one-to-one mapping between clusters and genres.

In order to quantify these differences, we use clustering evaluation metrics to compare automatically learned clusters with the manual genre annotation on the UN2000 corpus. Since clusters capture other distinctions than genres by design, note that we are repurposing the evaluation metrics to represent the distance between clusters and genres, rather than the quality of the clusters. We compute external clustering evaluation metrics as defined in [17]. Purity simply maps each cluster to its most frequent genre, and represents the percentage of correctly assigned documents. The remaining metrics view clustering as a series of decisions about each pair of documents. A decision is correct if two documents assigned to the same cluster are also assigned to the same genre. Rand Index represents the percentage of decisions that are correct (i.e. accuracy), while Precision/Recall/F-Measure separate the impact of false positives and false negatives. The relatively high purity and Rand Index scores reported in Table VIII can be explained by the imbalanced genre distribution which results in assigning the most frequent genre to most clusters. The lower precision and recall scores show that the clusters yield distinctions that are very different from genres.

This yields interesting differences between genre-based and cluster-based translation quality. While SMT yields much better BLEU scores than TM for all but one genre, the difference in BLEU scores between TM and SMT within clusters is more varied. In particular, for the 3 clusters that yield the highest BLEU scores, the TM does better than the SMT.

One key difference is that clusters are informed by the training documents since they are learned using the SMT/TM training corpus. In contrast the genres are defined independently from the training corpus.

#### C. Impact on translation workflow

These results seem to suggest a possible strategy for applying CAT tools, which could go as:

- For clusters 1, 3 and 5, the high BLEU scores obtained with SMT and the even higher scores obtained with TM suggest that a combination of TM and SMT [5] should provide results that need little or no post-edition (depending on the end user’s quality requirement);
- Clusters 2, 6, 7 and 8 should be suitable for SMT with human postediting. These clusters get lower BLEU scores, a higher percentage of documents in the worst BLEU quartiles and a bigger difference between SMT and TM performance.

Taken together, this suggests that SMT can yield reasonable translation quality for some documents, but that it cannot be trusted for all documents without human verification.

### VIII. CONCLUSION

We showed that translation quality for heterogeneous documents from the same UN source varies widely across genres and topic-based clusters, for two very different computer-aided translation tools: translation memories and statistical machine translation. This suggests that when translating large heterogeneous text collections such as those built from social media data, different translation strategies should be followed depending on the genres or clusters of documents to translate. SMT and TM can be used in combination for a subset of the documents, but both genres and cluster distinctions identify documents where translations probably require human post-edition in order to be useful.

This analysis departs from previous work in machine translation evaluation which focuses on comparing the performance of different systems on a fixed test set [1, for instance]. In contrast we have compared the performance of a single TM and a single SMT across different types of documents.

We plan to extend this work along several directions. First, we would like to further investigate unsupervised document clustering, and study whether topic-based clusters and genre distinctions provide complementary information for predicting translation quality. Second, we will leverage the multilingual nature of the UN corpus and evaluate whether our findings on Chinese-English translation hold for translation between other languages such as Arabic-English. Third, we will study whether genre and cluster corpus structure can be leveraged to improve the performance of CAT tools for specific document types, as in domain adaptation [18], [19]. Finally, we would like to experiment with document routing in actual translation workflows involving human translators and post-editors for translation tasks representative of the needs of their clients.

### ACKNOWLEDGMENT

The authors would like to thank John S. White and Timothy B. Allison for making data available for this project; Massih-Reza Amini, Michel Simard, and the Portage team at NRC for helpful comments and suggestions; and the three anonymous reviewers for their insightful feedback.

### REFERENCES

[1] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, July 2011, pp. 22–64.

[2] "NIST open machine translation (OpenMT) evaluation," <http://www.itl.nist.gov/iad/mig/tests/mt/>, last visited April 23, 2012.

[3] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of Machine Translation Summit X*, 2005.

[4] A. Finn and N. Kushmerick, "Learning to classify documents according to genre," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1506–1518, 2006. [Online]. Available: <http://dx.doi.org/10.1002/asi.20427>

[5] M. Simard and P. Isabelle, "Phrase-based machine translation in a computer-assisted translation environment," in *Proceedings of the twelfth Machine Translation Summit*, Ottawa, Ontario, Canada, August 2009, pp. 120–127.

[6] C. Lin and F. Och, "Orange: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation," in *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004, pp. 501–es.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[8] G. Foster, B. Chen, E. Joanis, H. Johnson, R. Kuhn, and S. Larkin, "PORTAGE in the NIST 2009 MT Evaluation," NRC-CNRC, Tech. Rep., 2009.

[9] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2009.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, March 2003. [Online]. Available: <http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993>

[11] M. Steyvers and T. Griffiths, "Probabilistic topic models," In *Handbook of Latent Semantic Analysis*, Hillsdale, NJ, 2007.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[13] Y.-M. Kim, "Document clustering in a learned concept space," Ph.D. dissertation, University Pierre and Marie Curie, 2010.

[14] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan, "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation," in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics/MATR*, Uppsala, Sweden, July 2010, pp. 17–53, revised August 2010.

[15] S. Vogel, H. Ney, and C. Tillmann, "HMM-based Word Alignment in Statistical Machine Translation," in *Proceedings of COLING'96*. Copenhagen, Denmark, 1996, pp. 836–841.

[16] R. Soricut and A. Echihiabi, "TrustRank: Inducing Trust in Automatic Translations via Ranking," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 612–621. [Online]. Available: <http://www.aclweb.org/anthology/P10-1063>

[17] C. D. Manning, P. Raghavan, and H. Schuetze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[18] G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 128–135. [Online]. Available: <http://www.aclweb.org/anthology/W/W07/W07-0717>

[19] G. Foster, C. Goutte, and R. Kuhn, "Discriminative instance weighting for domain adaptation in statistical machine translation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, October 2010, pp. 451–459. [Online]. Available: <http://www.aclweb.org/anthology/D10-1044>