Detecting Content-Heavy Sentences: A Cross-Language Case Study

Junyi Jessy Li
University of Pennsylvania
ljunyi@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract

The information conveyed by some sentences would be more easily understood by a reader if it were expressed in multiple sentences. We call such sentences content heavy: these are possibly grammatical but difficult to comprehend, cumbersome sentences. In this paper we introduce the task of detecting content-heavy sentences in cross-lingual context. Specifically we develop methods to identify sentences in Chinese for which English speakers would prefer translations consisting of more than one sentence. We base our analysis and definitions on evidence from multiple human translations and reader preferences on flow and understandability. We show that machine translation quality when translating content heavy sentences is markedly worse than overall quality and that this type of sentence are fairly common in Chinese news. We demonstrate that sentence length and punctuation usage in Chinese are not sufficient clues for accurately detecting heavy sentences and present a richer classification model that accurately identifies these sentences.

1 Introduction

To generate text, people and machines need to decide how to package the content they wish to express into clauses and sentences. There are multiple possible renderings of the same information, with varying degrees of ease of comprehension, compactness and naturalness. Some sentences, even though they are grammatical, would be more accessible to a reader if expressed in multiple sentences. We call such sentences *content heavy* sentences, or *heavy sentences* for brevity.

In the established areas of language research, text simplification and sentence planning in dialog and generation systems are clearly tasks in which identification of content-heavy sentences is of great importance. In this paper we introduce a novel flavor of the task in the cross-lingual setting, which in the long term may guide improvements in machine translation. We seek to identify sentences in Chinese that would result in heavy sentences in English if translated to a single sentence.

Example I in Table 1 shows a Chinese sentence and its two English translations A and B. Translator A used three English sentences to express all the information. Translator B, on the other hand, used a single sentence, which most readers would find more difficult to read. Example II illustrates a case where a translator would be hard pressed to convey all the content in a sentence in Chinese into a single grammatical English sentence.

Here we provide an operational characterization of content-heavy sentences in the context of Chinese-English translation. Instead of establishing guidelines for standalone annotation, we repurpose datasets developed for evaluation of machine translation consisting of multiple reference translations for each Chinese sentence. In this cross-lingual analysis, sentences in Chinese are considered content-heavy if their content would be more felicitously expressed in multiple sentences in English.

We first show that, with respect to English, content-heavy Chinese sentences are common. A fifth to a quarter of the sentences in the Chinese news data that we analyze are translated to multiple sentences in English. Moreover our experiments with reader preference indicate that for these sentences, readers strongly prefer multi-sentence translation to a single-sentence translation (§ 4.1). We also compare the difference in machine translation quality for heavy sentences and find that it is considerably lower than overall system performance (§ 4.2). Then we study the connection between heavy sentences and the factors

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1283–1293,
Lisbon, Portugal, 17-21 September 2015.

[Example I] 虽然菲军方在南部的巴西兰岛上部署了5000多兵力，并在美军的帮助下围剿阿布沙耶夫分子，但迄今收效不大。*Although the Philippine army on the southern Basilan island deployed over 5,000 troops, and with the US army's help are hunting down ASG members, but so far achieved little.* **[A]** The Philippine army has already deployed over 5 thousand soldiers on the southern island of Basilan. With the help of U.S. army, these soldiers are searching and suppressing members of Abu Sayyaf. However, there is not much achievement this far. **[B]** The Philippine military has stationed over 5,000 troops on Basilan Island in the southern Philippines and also tried to hunt down ASG members with the help of the United States, yet so far it has little success.	**[Example II]** 端粒是染色体末端的结构，随着细胞老化和失去分裂能力，端粒会逐渐缩短长度，换言之，端粒愈长显示细胞老化愈慢。*Telomeres are chromosome ends structures, with cell aging and losing division ability, telomeres will gradually decrease length, in other words, telomeres the longer shows cell aging the slower.* **[A]** Telomeres are structures at the ends of chromosomes, which gradually reduce in length with the aging of the cells and their loss of the ability to divide. In other words, longer telomeres indicate the slower aging of the cells. **[B]** Telomeres are the physical ends of chromosomes. As cells age and lose the ability to divide, the telomeres shrink gradually. That is to say, longer telomeres indicate that cells are aging more slowly.

Table 1: Examples of Chinese sentences expressed in multiple English sentences.

used in prior work to split a Chinese sentence into multiple sentences, showing that they do not fully determine the empirically defined content-heavy status (§ 5). Finally, we present an effective system to automatically identify content-heavy sentences in Chinese (§ 6, 7, 8).

2 Related work

The need for identifying content-heavy sentences arises in many specialized domains, including dialog systems, machine translation, text simplification and Chinese language processing but it is usually addressed in an implicit or application specific way. In contrast, we focus on identifying heavy sentences as a standalone task, providing a unifying view of the seemingly disparate strands of prior work. We now overview the literature which motivated our work.

Sentence planning. In text generation, a sentence planner produces linguistic realizations of a list of propositions (Rambow and Korelsky, 1992). One subtask is to decide whether to package the same content into one or more sentences. In the example below (Pan and Shaw, 2005), the multi-sentence expression B is much easier to process:

[A] This is a 1 million dollar 3 bedroom, 2 bathroom, 2000 square foot colonial with 2 acre of land, 2 car garage, annual taxes 8000 dollars in Armonk and in the Byram Hills school district.

[B] This is a 3 bedroom, 2 bathroom, 2000 square foot colonial located in Armonk with 2 acres of land. The asking price is 1 million dollar and the annual taxes are 8000 dollars. The house is located in the Byram Hills School District.

Identifying sentence [A] as heavy would be useful in selecting the best realization.

A crucial difference between our task and its counterpart in sentence planning is that traditional text generation systems have access to rich semantic information about the type of propositions the system needs to convey, while in our task we have access only to Chinese text. In some dialog systems, content selection is treated as an optimization problem, balancing the placement of full-stops and the insertion or deletion of propositions with the similarity of the resulting output and an existing corpus of acceptable productions (Pan and Shaw, 2005). Others formulate the problem as a supervised ranking task, in which different possible content realizations are generated, including variation in the number of sentences (Walker et al., 2001; Stent et al., 2004). With the introduction of the concept of content-heavy sentences, we can envision dialog systems addressing the sentence realization task in two steps, first predicting if the semantic content will require multiple sentences, then having different rankers for expressing the content in one or multiple sentences. In that case the ranker will need to capture only sentence-level information and the discourse-level decision to use multiple sentences will be treated separately.

Text simplification. "Text simplification, defined narrowly, is the process of reducing the linguistic complexity of a text, while still retaining the original information content and meaning" (Siddharthan, 2014). An important aspect of simplification is syntactic transformation in which sentences deemed difficult are re-written as multiple sentences (Chandrasekar et al., 1996; Aluísio et al., 2008). Our task may be viewed as identifying sentences in one language that will require simplification when translated, for the benefit of the speakers of the target language. In rule-based simplification systems, splitting is performed al-

ways when a given syntactic construction such as relative clause, apposition or discourse connective are detected (Chandrasekar et al., 1996; Siddharthan, 2006; De Belder and Moens, 2010). Most recently, text simplification has been addressed as a monolingual machine translation task from complex to simple language (Specia, 2010; Coster and Kauchak, 2011; Wubben et al., 2012). However simplification by repackaging the content into multiple sentences is not naturally compatible with the standard view of statistical MT in which a system is expected to produce a single output sentence for a single input sentence. Some of the recent systems using MT techniques separately model the need for sentence splitting (Zhu et al., 2010; Woodsend and Lapata, 2011; Narayan and Gardent, 2014). Identifying heavy sentences in simplification is equivalent to identifying sentences that require syntactic simplification.

Sentence structure and MT. Prior work in machine translation has discussed the existence of sentences in Chinese which would result in a poor translation if translated in one sentence in English. The main factors proposed to characterize such problematic sentences are sentence length (Xu and Tan, 1996) and the presence of given syntactic constructions (Xu et al., 2005; Yin et al., 2007; Jin and Liu, 2010). Mishra et al. (2014) used rules involving similar factors to distinguish sentences in Hindi that need simplification prior to translation.

In each of these approaches, the identified sentences are segmented into smaller units. Similar to work in text simplification, the simplification rules are applied to all sentences meeting certain criteria, normally to all sentences longer than a predefined threshold or where certain conjunctions or coordinations are present. In contrast, the model we propose here can be used to predict when segmentation is at all necessary.

Our approach to the problem is more compatible with the empirical evidence we presented in our prior work (Li et al., 2014) where we analyzed the output of Chinese to English machine translation and found that there is no correlation between sentence length and MT quality. Rather we showed that the quality of translation was markedly inferior, compared to overall translation quality, for sentences that were translated into multiple English sentences. This prior work was carried over a dataset containing a single reference translation for each Chinese sentence. In the work presented in this paper, we strengthen our findings by examining multiple reference translations for each Chinese sentence. We define heavy sentences based on agreement of translator choices and reader preferences.

Commas in Chinese. Often a comma in a sentence can be felicitously replaced by a full stop. Such commas offer a straightforward way to split a long sentence into multiple shorter ones by replacing the comma with a full stop. Monolingual text simplification systems often try to identify such commas. They are particularly common in Chinese and replacing them with full stops leads to improvements in the accuracy of syntactic parsing (Jin et al., 2004; Li et al., 2005). Moreover, existing syntactically parsed corpora conveniently provide numerous examples of these full-stop commas, and thus training data for systems to identify them (Xue and Yang, 2011; Yang and Xue, 2012). In this paper, we systematically study the relationship between the presence of full-stop commas in the sentence and whether it is content-heavy for Chinese to English translation.

3 Data

In this work we use three news datasets: the newswire portion of the NIST 2012 Open Machine Translation Evaluation (OpenMT) (Group, 2013), Multiple-Translation Chinese (MTC) parts 1-4 (Huang et al., 2002; Huang et al., 2003; Ma, 2004; Ma, 2006), and the Chinese Treebank (Xue et al., 2005). In OpenMT and MTC, multiple reference translations in English are available for each Chinese segment (sentence).

To study the relationship between content-heavy sentences and reader preference for multi-sentence translations (§ 4.1), we use OpenMT (688 segments) and MTC parts 2-4 (2,439 segments), both of which provide four English translations for each Chinese segment. This analysis forms the basis for labeling heavy sentences for supervised training and evaluation (§ 5, 6, 7).

The Chinese Treebank (CTB) has been used in prior work as data for identifying full-stop commas. Moreover, 52 documents in MTC part 1 were drawn from the CTB. The intersection of the two datasets allows us to directly analyze the relationship between heavy sentences and full-stop commas in Chinese (§ 5). Furthermore we use this intersection as test set to identify heavy sentences so we can directly compare with models developed

for comma disambiguation. To be consistent with the rest of the MTC data, we use 4 out of the 11 translators in part 1 in these experiments.[1]

Our model for Chinese full-stop comma recognition is trained following the features and training sets specified in Xue and Yang (2011)[2], excluding the overlapping MTC/CTB documents mentioned above. There are 12,291 sentences in training that contain at least one comma. A classifier for detecting heavy sentences is trained on OpenMT and MTC (excluding the test set). A quick inspection of both datasets reveals that Chinese sentences without a comma were never translated into multiple sentences by more than one translator. Therefore in our experiments we consider only sentences that contain at least one comma. There are 301 test sentences, 511 training sentences in OpenMT and 2418 in MTC. Sentences are processed by the Stanford NLP packages[3]. CTB gold-standard parses are used to obtain full-stop commas and to train comma disambiguation models.

4 Content-heavy sentences: definition

In this section we provide an operational definition for which sentences should be considered content-heavy, based on the choices made by translators and the fluency preferences of readers when a sentence is translated into a single or multiple sentences. We further demonstrate the difference in machine translation quality when translating content-heavy sentences compared to other sentences.

4.1 Content-heaviness and multi-sentence translations

First we quantify how often translators choose to translate a Chinese sentence into multiple English sentences. Content-heavy Chinese sentences are those for which there is a strong preference to produce multiple sentences when translating to English (at the end of the section we present specific criteria).

Obviously, splitting a sentence into multiple ones is often possible but is not necessarily preferred. In Table 2, we show in the "%data" columns the percentage of source sentences split in translation by 0, 1, 2, 3 and all 4 translators. For about 20% of segments in OpenMT and 15% in MTC, at least three of the translators produce a multi-sentence translation, a rate high enough to warrant closer inspection of the problem.

	OpenMT		MTC	
#ref multi	%data	%best multi	%data	%best multi
0	65.4	0	58.9	0
1	7.4	23.5	20.4	20.1
2	7.0	66.7	8.3	56.7
3	9.2	88.9	7.9	89.6
4	11.0	100	4.6	100

Table 2: Percentage of sentences for which a given number of translators prefer to use multiple sentences in English, along with percentage of times a multi-sentence translation was selected as most fluent and comprehensible by readers.

Next, we conduct a study to find out what level of translator agreement leads to strong reader preference for the same information to be presented in multiple sentences.

For each Chinese segment with one, two or three multi-sentence reference translations, we ask five annotators on Mechanical Turk to rank the reference translations according to their general flow and understandability. The annotators saw only the four randomly ordered English translations and were not shown the Chinese original, with the following instruction:

Below are 1-2 sentence snippets that describe the same content. Some are more readable and easier to understand than others. Your task is to rank them from the best to worst in terms of wording or flow (organization). There can be ties, but you have to pick one that is the best.

We obtain reader preference for each segment in the following manner: for each annotator, we take the highest ranked translation and check whether it consists of multiple sentences. In this way we have five binary indicators. We say readers prefer a sentence to have a multi-sentence translation in terms of flow and comprehensibility if the majority of these five indicators are positive.

In the "%best multi" columns of Table 2, we tabulate the percentage of segments with majority preference for multi-sentence translation, stratified by the number of translators who split the content. Obviously the more multi-sentence translations there are, the higher the probability that the readers will select one as the best translation. We are interested in knowing for which conditions the

[1] We did not use translator IDs as parameters in any of our systems.

[2] Document IDs 41-325, 400-454, 500-554, 590-596, 600-885, 900, 1001-1078, 1100-1151.

[3] The Stanford segmenter (Tseng et al., 2005), parser (Levy and Manning, 2003) and the CoreNLP package (Manning et al., 2014)

Criteria	%data(Y)	Y	N	Δbleu
heavy	27.2	15.34	19.24	**3.9**

Table 3: Percentage of data for heavy sentences along with BLEU scores for heavy and non-heavy sentences and their difference.

preference for multi-sentence translation exceeds the probability of randomly picking one.

When only one (out of four) translations is multi-sentence, the best translations chosen by the majority of readers contain multiple sentences less often than in random selection from the available translations. When two out of the four reference translations are multi-sentence, the reader preference towards them beats chance by a good margin. The difference between chance selection and reader preference for multiple sentences grows steadily with the number of reference translations that split the content. These data suggest that when at least two translators perform a multi-sentence translation, breaking down information in the source sentence impacts the quality of the translation.

Hence we define content-heavy sentences in Chinese to be those for which at least two out of four reference translations consist of multiple sentences.

4.2 A challenge for MT

We now quantitatively show that heavy sentences are particularly problematic for machine translation. We collect translations for each segment in OpenMT and MTC from the Bing Translator. We split the sentences into two groups, heavy and other, according to the gold standard label explained in the previous section. We then compare the BLEU score for sentences in a respective group, where each group is in turn used as a test set. The difference in BLEU scores (Δbleu) is a strong indicator whether these sentences are challenging for MT systems.

In Table 3 we show the BLEU scores and Δbleu for sentences that are heavy (Y) and non-heavy (N). Also included in the table is the percentage of heavy sentences in all the data.

Translations for heavy sentences received a BLEU score that is 3.9 points lower than those that are not. This clearly illustrates the challenge and potential for improvement for MT systems posed by content-heavy sentences. Therefore the ability to reliably recognize them provides a first step towards developing a better translation approach for such sentences.

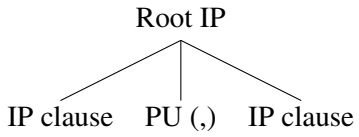

Figure 1: Coordinating IP structure at the root.

heavy	fs-comma	No fs-comma
N	19	180
Y	40	62

Table 4: Count of heavy and non-heavy sentences with and without full-stop commas.

5 Comma usage and heavy sentences

In Chinese, commas can sometimes act as sentence boundaries, similar to the function of an English period. In Xue and Yang (2011), the authors showed that these full-stop commas can be identified in the constituent parse tree as coordinating IPs at the root level, shown in Figure 1. Fancellu and Webber (2014) demonstrated that it is beneficial to split sentences containing negation on these types of commas, translate the resulting shorter sentences separately, then stitch the resulting translations together. They report that this approach prevented movement of negation particles beyond their scope. Here we study the degree to which the content-heavy status of a sentence is explained by the presence of a full-stop comma in the sentence. We show that they are interrelated but not equivalent.

Corpus analysis. First we study how often a heavy sentence contains a full-stop comma and vice versa, using the overlapping MTC/CTB documents. We show in Table 4 the number of heavy and non-heavy sentences with and without full-stop commas[4]. When there is a full-stop comma in the sentence, there is a higher chance that the sentence is content-heavy. Yet of the 102 heavy sentences in this data, fewer than 40% contain full-stop commas; of the 242 sentences without full-stop commas, more than a quarter are heavy. Therefore, although comma usage in the Chinese sentence may provide clues for detecting content heaviness, the two phenomena are not equivalent

[4]For the study we exclude sentences without a comma. A χ^2 test for the strength of association between the presence of full stop commas and heavy sentence status shows high significance.

and heavy sentences are not fully explained by the presence of full-stop commas.

Learning with full-stop commas. Here we evaluate the usefulness of using full-stop commas as training data to predict whether a sentence is content-heavy. From the analysis presented above we know that the two tasks are not equivalent. Nevertheless we would like to test directly if the Chinese Treebank—the large (but noisy for the task at hand) data available for comma function disambiguation—would lead to better results than learning on the cleaner but much smaller datasets for which multiple translations are available.

We use logistic regression as our classification model[5]. The performance of identifying heavy sentences on the MTC/CTB overlapping test set is compared using the following methods:

[Parallel] A classifier is trained using four English translations for each Chinese sentence (OpenMT and MTC training set). Following the definition in Section 4.1, content-heavy sentences are those translated into multiple English sentences by two or more translators.

[Oracle comma] A test sentence is assigned to class "heavy" if there is a full-stop comma in its corresponding gold standard parse tree.

[Predicted comma] We train a comma disambiguation system on CTB to predict if a comma is a full-stop comma. In testing, a sentence is marked "heavy" if it contains a predicted full-stop comma.

Features. We reimplemented the per-comma features used in Xue and Yang (2011)[6]. As in their best performing system, features are extracted from gold-standard parse trees during training and from automatic parsing during testing. These include: words and part-of-speech tags immediately before and after the comma; left- and right-sibling node labels of the parent of the comma; ordered ancestor node labels above the comma; punctuation tokens ordered from left to right of the sentence; whether the comma has a coordinating IP structure; whether the comma's parent is a child of the root of the tree; whether there is a subordination before the comma; whether the difference in number of words before and after the comma is greater than or equal to seven.

[5]We use the Liblinear package (Fan et al., 2008).

[6]For *predicted comma*, our reimplementation of Xue and Yang (2011) gave practically identical results to those reported in the original paper on the test set that they used.

Training	A	P	R
parallel	75.75	69.86	50
oracle comma	73.09	67.8	39.2
predicted comma	74.42	66.67	49.02

Table 5: Performance for identify heavy sentences using multiple reference data (parallel) vs. full-stop comma oracle labels (oracle comma) and predicted full-stop commas (predicted comma).

For *parallel*, feature values are accumulated from all the commas in the sentence. For binary features, we use an *or* operation on the feature values for each individual comma.

Results and comparison. In Table 5, we show the accuracy, precision and recall for identifying content-heavy sentences using the three methods described above. We do not include the majority baseline here because it assumes no sentences are content heavy.

Interestingly, the system using oracle information in each test sentence for full-stop commas performs the worst. The system trained to identify full-stop commas outperform the oracle system with about 10% better in recall and less than 1% lower in precision. This finding strongly suggests that the features used for learning capture certain characteristics of heavy sentences even with non-ideal training labels. The best performance is obtained learning directly on parallel corpora with multiple reference translations. Note that we try to provide the best possible setting for full-stop comma prediction, using much more training data, gold-standard parses, same-domain training and testing, as well as the reimplementation of state-of-the-art system. These settings allow us to conservatively interpret the results listed here, which confirm that content-heaviness is different from using a full-stop comma in the Chinese sentence. It is more advantageous—leading to higher precision and overall accuracy—to learn from data where translators encode their interpretation in the form of multi-sentence translations.

6 Features to characterize content-heavy sentences

In this section, we experiment with a wide range of features from the sentence string, part-of-speech tags and dependency parse trees.

Baseline. Intuitively, sentence length can be an indication of too much content that needs to be

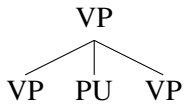

Figure 2: Multiple VP structures

repackaged into multiple sentences. Therefore as our baseline we train a decision tree using the number of words[7] in a Chinese sentence.

Sentence structure cues. We collect potential signals for structural complexity: punctuation, conjunctions, prepositional phrases and relative clauses. As features we count the number of commas, conjunction, preposition and postposition part-of-speech tags. In Chinese "DE" often marks prepositional phrases or relative clauses among other functions (Chang et al., 2009a). Here we include a simple count the number of "DEG" tags in the sentence.

Dependencies. Dependency grammar captures both syntactic and semantic relationship between words and are shown to improve reordering in MT (Chang et al., 2009b). To account for such relational information we include two feature classes: the percentage of each dependency type and the typed dependency pairs themselves. For the latter we use the universal part-of-speech tags (Petrov et al., 2012) for each word rather than the word itself to avoid too detailed and sparse representations. For example, the relation *dobj(处理/handle,* 事情*/matter)* becomes feature *dobj(verb, noun)*.

Furthermore, we use dependency trees to extract four features for potentially complex constructions. First, we indicate the presence of noun phrases with heavy modifiers on the left. These are frequently used in Chinese and would require a relative clause or an additional sentence in English. Specifically we record the maximum number of dependents for the nouns in the sentence. The second type of construction is the use of serial verb phrases, illustrated in Figure 2. We record the number of dependents of the head verb of the sentence. The third feature class is the typed dependencies (over universal POS tags) whose edge crosses a comma. Finally, we also record the maximum number of dependents in the sentence to capture the general phrasal complexity in the sentence.

[7]obtained using the Stanford Chinese Word Segmenter (Tseng et al., 2005)

Features	Training	A	P	R
baseline	MTC+OpenMT	71.43	73.5	24.5
full set	OpenMT	76.41	66.67	**60.78**
full set	MTC	78.41	74.03	55.9
full set	MTC+OpenMT	**80.73**	**79.73**	57.84

Table 6: Accuracy, precision and recall (for the content heavy class) of binary classification using proposed features to identify content-heavy sentences.

Parts-of-speech. POS information captures numerous aspects of the sentence such as the frequency of different classes of words used and the transition between them. Historically they are also shown to be helpful for phrase boundary detection (Taylor and Black, 1998). Here, we first convert all Chinese POS tags into their corresponding universal tags. We then use the percentage of each tag and tag bigram as two feature classes. To capture the transition of each phrase and clause in the sentence, we construct functional POS trigrams for each sentence by removing all nouns, verbs, adjectives, adverbs, numbers and pronouns in the sentence. Percentages of these sequences are used as feature values.

Comma disambiguation features. We also incorporate most of the features proposed by Xue and Yang (2011), aggregated in the same way as the *parallel* method (cf. Section 5). These include: POS tags immediately before and after the comma; left- and right-sibling node labels of the parent of the comma; the punctuation tokens ordered from left to right in the sentence, whether the comma has a coordinating IP structure; whether the comma's parent is a child of the root of the tree; whether there is a subordination before the comma; whether the difference in number of words before and after the comma is greater than or equal to seven.

7 Results

7.1 Recognizing content-heavy sentences

We train a logistic regression model as in the *parallel* method in Section 5 using features illustrated above. In Table 6, we show the performance of detecting heavy sentences using four systems: the baseline system using the number of words in the sentence and three systems using our full feature set, trained on MTC, OpenMT and both.

The baseline performance is characterized by a remarkably poor recall. It becomes appar-

#ref multi	≥ 0	≥ 1	≥ 2	≥ 3	4
#seg	301	187	102	58	25
precision	79.73	84.29	100	100	100
recall	57.85	57.84	57.84	68.98	76
posterior	0.29	0.40	0.53	0.61	0.67

Table 7: Number of segments, precision, recall and posterior probability (for the content-heavy class) for examples where at least 0-4 translators split the sentence.

ent that length alone cannot characterize content-heaviness. On the other hand, using the full feature set achieves an accuracy of above 80%, a precision close to 80% and a recall about 58%. The improvement in precision and recall over using oracle full-stop commas (Table 5) are about 12% and 19%. When compared with using features tuned for comma disambiguation from Xue and Yang (2011) (Table 5), our full feature set achieved a 5% increase in accuracy, about 10% increase in precision and 8% increase in recall.

We also demonstrate the usefulness of having more multi-reference translation data by comparing training using MTC and OpenMT individually and both. Remarkably, using only the very small dataset of OpenMT is sufficient to produce a predictor that is more accurate than all of the methods listed in Section 5. Adding these examples to MTC drastically improves precision by more than 13% with a less than 3% drop on recall.

Finally, we consider the portions of our test set for which at least n translators provided a multi-sentence translation (n ranges from 0 to 4). In Table 7 we show the respective precision, recall and the average posterior probability from the classifier for marking a sentence as content-heavy. There is a clear trend that the classifier is more confident and has higher precision for sentences where more translators produce multi-sentence translations. Although the model is not highly confident in all groups, the precision of the predictions are remarkably high. Miss rate also decreases when more translators translate the source into multiple sentences.

7.2 Post-hoc feature analysis

Here we identify which of the feature classes from our full set are most helpful by performing forward feature selection: in each iteration, the feature class that improves accuracy the most is selected. The process is repeated until none of the remaining feature classes lead to improvement when added to the model evaluated at the previous iteration. We use our test data as the evaluation set for forward selection, but we do so only to evaluate features, not to modify our system.

Five feature classes are selected using this greedy procedure. The first selected class is the typed dependencies over universal POS tags. Remarkably, this single feature class achieves 76.6% accuracy, a number already reasonably high and better than features used in Xue and Yang (2011). The second feature added is whether there is a comma of coordinating IP structure in the automatic parse tree of the sentence. It gives a further 1.7% increase in accuracy, showing that the comma structure provide useful information as features for detecting heavy sentences. Note that this feature does not represent full stop commas, i.e., it does not record whether the comma is under the root level of the parse tree. The next selected class is typed dependencies over universal POS tags that have an edge across commas in the sentence, with an 1% increase in accuracy. The fourth feature selected is the number of prepositions and postposition POS tags in the sentence, improving the accuracy about 1%. Finally, part-of-speech tags before each comma are added, with a 0.3% improvement of accuracy.

The results from forward selection analysis reveal that the dependency structure of a sentence captures the most helpful information for heavy sentence identification. The interplay between punctuation and phrase structure gives further important enhancements to the model. The final accuracy, precision and recall after forward selection are 0.804, 0.8209, 0.5392, respectively. This overall performance shows that forward selection yields a sub-optimal feature set, suggesting that the other features are also informative.

8 A challenge for MT: revisited

It is important to know whether a predictor for content-heavy sentences is good at identifying challenging sentences for applications such as machine translation. Here, we would like to revisit Section 4.2 and see if *predicted* heavy sentences are harder to translate.

For all the source sentences in OpenMT and MTC, we compare five criteria for dividing the test data in two subsets: whether the sentence contains a full-stop comma or not; whether the sentence is longer than the baseline decision tree threshold

Criteria	%data(Y)	Y	N	Δbleu
fs-comma	21.6	16.01	18.43	2.42
length threshold	8.6	15.38	18.3	2.92
pred-heavy (0.5)	22.72	15.81	18.77	2.96
pred-heavy (0.55)	19.72	15.47	18.76	3.29
pred-heavy (0.6)	16.67	14.95	18.77	3.82
oracle heavy	27.4	15.34	19.24	3.9

Table 8: Data portion, BLEU scores and differences for sentences with/without a full-stop comma, are/are not longer than the length threshold, are/are not content heavy.

(47 words) or not; whether the sentence is predicted to be content-heavy with posterior probability threshold of 0.5, 0.55 and 0.6. Predictions for the training portion is obtained using 10-fold cross-validation. In the same manner as Table 3, Table 8 shows the percentage of data that satisfies each criterion, BLEU scores of Bing translations for sentences that satisfy a criterion and those that do not, as well as the difference of BLEU between the two subsets (Δbleu). As reference we also include numbers listed in Table 3 using oracle content-heavy labels.

First, notice that regardless of the posterior probability threshold, the numbers of sentences predicted to be content-heavy are much larger than that using the length cutoff. These sentences are also collectively translated much worse than the sentences in the other subset. Sentences that contain a predicted full-stop comma are also harder to translate, but show smaller difference in BLEU than when sentence heaviness or length are used as separation criterion. As the posterior probability threshold goes up and the classifier becomes more confident when it identifies heavy sentences, there is a clear trend that system translations for these sentences become worse. These BLEU score comparisons indicate that our proposed model identifies sentences that pose a challenge for MT systems.

9 Conclusion and future work

In this work, we propose a cross-lingual task of detecting content-heavy sentences in Chinese, which are best translated into multiple sentences in English. We show that for such sentences, a multi-sentence translation is preferred by readers in terms of flow and understandability. Content-heavy sentences defined in this manner present practical challenges for MT systems. We further demonstrate that these sentences are not fully explained by sentence length or syntactically defined full-stop commas in Chinese. We propose a classification model using a rich set of features that effectively identify these sentences.

The findings in this paper point out a definite issue in different languages currently under-investigated in text-to-text generation systems. One possible way to improve MT systems is to incorporate sentence simplification before translation (Mishra et al., 2014). Future work could use our proposed model to detect heavy sentences that needs such pre-processing. Our findings can also inspire informative features for sentence quality estimation, in which the task is to predict the sentence-level fluency (Beck et al., 2014). We have shown that heavy Chinese sentences are likely to lead to hard to read, disfluent sentences in English. Another important future direction lies in text simplification. In our inspection of parallel Wikipedia/Simple Wikipedia data (Kauchak, 2013), around 23.6% of the aligned sentences involve a single sentence on one side and multiple sentences on another. A similar analysis using ideas from this work can be useful in identifying sentences that needs simplification in the first place.

Acknowledgements

We would like to express our gratitude to Bonnie Webber for her detailed comments on earlier versions of this paper. Her encouragement and support were essential in seeing the work through to publication. We would also like the reviewers for their thoughtful suggestions which we have tried to incorporate in the final version. The work was partially supported by NSF CAREER grant IIS-0953445.

References

Sandra M. Aluísio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceedings of the Eighth ACM Symposium on Document Engineering (DocEng)*.

Daniel Beck, Kashif Shah, and Lucia Specia. 2014. SHEF-Lite 2.0: Sparse multi-task gaussian processes for translation quality estimation. In *Ninth Workshop on Statistical Machine Translation (WMT)*.

R. Chandrasekar, Christine Doran, and B. Srinivas.

1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Pi-Chuan Chang, Daniel Jurafsky, and Christopher D. Manning. 2009a. Disambiguating "DE" for Chinese-English machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WSMT)*.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009b. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST)*.

Will Coster and David Kauchak. 2011. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*.

Federico Fancellu and Bonnie Webber. 2014. Applying the semantics of negation to SMT through n-best list re-ranking. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

NIST Multimodal Information Group. 2013. NIST 2008-2012 Open Machine Translation (OpenMT) Progress Test Sets LDC2013T07. Web Download. In *Philadelphia: Linguistic Data Consortium*.

Shudong Huang, David Graff, and George Doddington. 2002. Multiple-Translation Chinese Corpus LDC2002T01. Web Download. In *Philadelphia: Linguistic Data Consortium*.

Shudong Huang, David Graff, Kevin Walker, David Miller, Xiaoyi Ma, Christopher Cieri, and George Doddington. 2003. Multiple-Translation Chinese (MTC) Part 2 LDC2003T17. Web Download. In *Philadelphia: Linguistic Data Consortium*.

Yaohong Jin and Zhiying Liu. 2010. Improving Chinese-English patent machine translation using sentence segmentation. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*.

Meixun Jin, Mi-Young Kim, Dongil Kim, and Jong-Hyeok Lee. 2004. Segmentation of Chinese long sentences using commas. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing (SIGHAN)*.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, January.

Xing Li, Chengqing Zong, and Rile Hu. 2005. A hierarchical parsing approach with punctuation processing for long Chinese sentences. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJNLP): Companion Volume including Posters/Demos and Tutorial Abstracts*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers*.

Xiaoyi Ma. 2004. Multiple-Translation Chinese (MTC) Part 3 LDC2004T07. Web Download. In *Philadelphia: Linguistic Data Consortium*.

Xiaoyi Ma. 2006. Multiple-Translation Chinese (MTC) Part 4 LDC2006T04. Web Download. In *Philadelphia: Linguistic Data Consortium*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*.

Kshitij Mishra, Ankush Soni, Rahul Sharma, and Dipti Sharma. 2014. Exploring the effects of sentence simplification on Hindi to English machine translation system. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA)*.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shimei Pan and James Shaw. 2005. Instance-based sentence boundary determination by optimization for natural language generation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*.

Owen Rambow and Tanya Korelsky. 1992. Applied text generation. In *Proceedings of the Third Conference on Applied Natural Language Processing (ANLP)*.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165:259–298.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*.

Amanda Stent, Rashmi Prassad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL)*.

Paul Taylor and Alan W. Black. 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*, 12:99–117.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Prroceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Marilyn Walker, Owen Rambow, and Monica Rogati. 2001. Spot: A trainable sentence planner. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Donghua Xu and Chew Lim Tan. 1996. Automatic alignment of English-Chinese bilingual texts of CNS news. In *Proceedings of International Conference on Chinese Computing*.

Jia Xu, Richard Zens, and Hermann Ney. 2005. Sentence segmentation using IBM word alignment model 1. In *In Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT)*.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT): Short Papers*.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238, June.

Yaqin Yang and Nianwen Xue. 2012. Chinese comma disambiguation for discourse analysis. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dapeng Yin, Fuji Ren, Peilin Jiang, and Shingo Kuroiwa. 2007. Chinese complex long sentences processing method for Chinese-Japanese machine translation. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*.