Who, What, When, Where, Why?
Comparing Multiple Approaches to the Cross-Lingual 5W Task

**Kristen Parton*, Kathleen R. McKeown*, Bob Coyne*, Mona T. Diab*,
Ralph Grishman†, Dilek Hakkani-Tür‡, Mary Harper§, Heng Ji•, Wei Yun Ma*,
Adam Meyers†, Sara Stolbach*, Ang Sun†, Gokhan Tur˚, Wei Xu† and Sibel Yaman‡**

*Columbia University
New York, NY, USA
{kristen, kathy, coyne, mdiab, ma, sara}@cs.columbia.edu

‡International Computer Science Institute
Berkeley, CA, USA
{dilek, sibel}@icsi.berkeley.edu

•City University of New York
New York, NY, USA
hengji@cs.qc.cuny.edu

†New York University
New York, NY, USA
{grishman, meyers, asun, xuwei}@cs.nyu.edu

§Human Lang. Tech. Ctr. of Excellence, Johns Hopkins and U. of Maryland, College Park
mharper@umd.edu

˚SRI International
Palo Alto, CA, USA
gokhan@speech.sri.com

Abstract

Cross-lingual tasks are especially difficult due to the compounding effect of errors in language processing and errors in machine translation (MT). In this paper, we present an error analysis of a new cross-lingual task: the 5W task, a sentence-level understanding task which seeks to return the English 5W's (Who, What, When, Where and Why) corresponding to a Chinese sentence. We analyze systems that we developed, identifying specific problems in language processing and MT that cause errors. The best cross-lingual 5W system was still 19% worse than the best monolingual 5W system, which shows that MT significantly degrades sentence-level understanding. Neither source-language nor target-language analysis was able to circumvent problems in MT, although each approach had advantages relative to the other. A detailed error analysis across multiple systems suggests directions for future research on the problem.

1 Introduction

In our increasingly global world, it is ever more likely for a mono-lingual speaker to require information that is only available in a foreign language document. Cross-lingual applications address this need by presenting information in the speaker's language even when it originally appeared in some other language, using machine translation (MT) in the process. In this paper, we present an evaluation and error analysis of a cross-lingual application that we developed for a government-sponsored evaluation, the *5W task*.

The *5W task* seeks to summarize the information in a natural language sentence by distilling it into the answers to the 5W questions: Who, What, When, Where and Why. To solve this problem, a number of different problems in NLP must be addressed: predicate identification, argument extraction, attachment disambiguation, location and time expression recognition, and (partial) semantic role labeling. In this paper, we address the *cross-lingual 5W task*: given a source-language sentence, return the 5W's translated (comprehensibly) into the target language. Success in this task requires a synergy of successful MT and answer selection.

The questions we address in this paper are:

- How much does machine translation (MT) degrade the performance of cross-lingual 5W systems, as compared to monolingual performance?
- Is it better to do source-language analysis and then translate, or do target-language analysis on MT?
- Which specific problems in language processing and/or MT cause errors in 5W answers?

In this evaluation, we compare several different approaches to the cross-lingual 5W task, two that work on the target language (English) and one that works in the source language (Chinese).

A central question for many cross-lingual applications is whether to process in the source language and then translate the result, or translate documents first and then process the translation. Depending on how errorful the translation is, results may be more accurate if models are developed for the source language. However, if there are more resources in the target language, then the translate-then-process approach may be more appropriate. We present a detailed analysis, both quantitative and qualitative, of how the approaches differ in performance.

We also compare system performance on human translation (which we term reference translations) and MT of the same data in order to determine how much MT degrades system performance. Finally, we do an in-depth analysis of the errors in our 5W approaches, both on the NLP side and the MT side. Our results provide explanations for why different approaches succeed, along with indications of where future effort should be spent.

2 Prior Work

The cross-lingual 5W task is closely related to cross-lingual information retrieval and cross-lingual question answering (Wang and Oard 2006; Mitamura et al. 2008). In these tasks, a system is presented a query or question in the target language and asked to return documents or answers from a corpus in the source language. Although MT may be used in solving this task, it is only used by the algorithms – the final evaluation is done in the source language. However, in many real-life situations, such as global business, international tourism, or intelligence work, users may not be able to read the source language. In these cases, users must rely on MT to understand the system response. (Parton et al. 2008) examine the case of "translingual" information retrieval, where evaluation is done on translated results in the target language. In cross-lingual information extraction (Sudo et al. 2004) the evaluation is also done on MT, but the goal is to learn knowledge from a large corpus, rather than analyzing individual sentences.

The 5W task is also closely related to Semantic Role Labeling (SRL), which aims to efficiently and effectively derive semantic information from text. SRL identifies predicates and their arguments in a sentence, and assigns roles to each argument. For example, in the sentence "I baked a cake yesterday.", the predicate "baked" has three arguments. "I" is the subject of the predicate, "a cake" is the object and "yesterday" is a temporal argument.

Since the release of large data resources annotated with relevant levels of semantic information, such as the FrameNet (Baker et al., 1998) and PropBank corpora (Kingsbury and Palmer, 2003), efficient approaches to SRL have been developed (Carreras and Marquez, 2005). Most approaches to the problem of SRL follow the Gildea and Jurafsky (2002) model. First, for a given predicate, the SRL system identifies its arguments' boundaries. Second, the Argument types are classified depending on an adopted lexical resource such as PropBank or FrameNet. Both steps are based on supervised learning over labeled gold standard data. A final step uses heuristics to resolve inconsistencies when applying both steps simultaneously to the test data.

Since many of the SRL resources are English, most of the SRL systems to date have been for English. There has been work in other languages such as German and Chinese (Erk 2006; Sun 2004; Xue and Palmer 2005). The systems for the other languages follow the successful models devised for English, e.g. (Gildea and Palmer, 2002; Chen and Rambow, 2003; Moschitti, 2004; Xue and Palmer, 2004; Haghighi et al., 2005).

3 The Chinese-English 5W Task

3.1 5W Task Description

We participated in the 5W task as part of the DARPA GALE (Global Autonomous Language Exploitation) project. The goal is to identify the 5W's (Who, What, When, Where and Why) for a complete sentence. The motivation for the 5W task is that, as their origin in journalism suggests, the 5W's cover the key information nuggets in a sentence. If a system can isolate these pieces of information successfully, then it can produce a précis of the basic meaning of the sentence. Note that this task differs from QA tasks, where "Who" and "What" usually refer to definition type questions. In this task, the 5W's refer to semantic roles within a sentence, as defined in Table 1.

In order to get all 5W's for a sentence correct, a system must identify a top-level predicate, extract the correct arguments, and resolve attachment ambiguity. In the case of multiple top-level predicates, any of the top-level predicates may be chosen. In the case of passive verbs, the Who is the agent (often expressed as a "by clause", or not stated), and the What should include the syntactic subject.

Answers are judged Correct[1] if they identify a correct null argument or correctly extract an argument that is present in the sentence. Answers are not penalized for including extra text, such as prepositional phrases or subordinate clauses, unless the extra text includes text from another answer or text from another top-level predicate. In sentence 2a in Table 2, returning "bought and cooked" for the What would be Incorrect. Similarly, returning "bought the fish at the market" for the What would also be Incorrect, since it contains the Where. Answers may also be judged Partial, meaning that only part of the answer was returned. For example, if the What contains the predicate but not the logical object, it is Partial.

Since each sentence may have multiple correct sets of 5W's, it is not straightforward to produce a gold-standard corpus for automatic evaluation. One would have to specify answers for each possible top-level predicate, as well as which parts of the sentence are optional and which are not allowed. This also makes creating training data for system development problematic. For example, in Table 2, the sentence in 2a and 2b is the same, but there are two possible sets of correct answers. Since we could not rely on a gold-standard corpus, we used manual annotation to judge our 5W system, described in section 5.

3.2 The Cross-Lingual 5W Task

In the cross-lingual 5W task, a system is given a sentence in the source language and asked to produce the 5W's in the target language. In this task, both machine translation (MT) and 5W extraction must succeed in order to produce correct answers. One motivation behind the cross-lingual 5W task is MT evaluation. Unlike word- or phrase-overlap measures such as BLEU, the 5W evaluation takes into account "concept" or "nugget" translation. Of course, only the top-level predicate and arguments are evaluated, so it is not a complete evaluation. But it seeks to get at the understandability of the MT output, rather than just n-gram overlap.

Translation exacerbates the problem of automatically evaluating 5W systems. Since translation introduces paraphrase, rewording and sentence restructuring, the 5W's may change from one translation of a sentence to another translation of the same sentence. In some cases, roles may swap. For example, in Table 2, sentences 1a and 1b could be valid translations of the same Chinese sentence. They contain the same information, but the 5W answers are different. Also, translations may produce answers that are textually similar to correct answers, but actually differ in meaning. These differences complicate processing in the source followed by translation.

Example: On Tuesday, President Obama met with French President Sarkozy in Paris to discuss the economic crisis.

W	Definition	Example answer
WHO	Logical subject of the top-level predicate in WHAT, or null.	President Obama
WHAT	One of the top-level predicates in the sentence, and the predicate's logical object.	met with French President Sarkozy
WHEN	ARGM-TMP of the top-level predicate in WHAT, or null.	On Tuesday
WHERE	ARGM-LOC of the top-level predicate in WHAT, or null.	in Paris
WHY	ARGM-CAU of the top-level predicate in WHAT, or null.	to discuss the economic crisis

Table 1. Definition of the 5W task, and 5W answers from the example sentence above.

4 5W System

We developed a 5W combination system that was based on five other 5W systems. We selected four of these different systems for evaluation: the final combined system (which was our submission for the official evaluation), two systems that did analysis in the target-language (English), and one system that did analysis in the source language (Chinese). In this section, we describe the individual systems that we evaluated, the combination strategy, the parsers that we tuned for the task, and the MT systems.

	Sentence	WHO	WHAT
1a	Mary bought a cake from Peter.	Mary	bought a cake
1b	Peter sold Mary a cake.	Peter	sold Mary
2a	I bought the fish at the market yesterday and cooked it today.	I	bought the fish [WHEN: yesterday]
2b	I bought the fish at the market yesterday and cooked it today.	I	cooked it [WHEN: today]

Table 2. Example 5W answers.

[1] The specific guidelines for determining correctness were formulated by BAE.

4.1 Latent Annotation Parser

For this work, we have re-implemented and enhanced the Berkeley parser (Petrov and Klein 2007) in several ways: (1) developed a new method to handle rare words in English and Chinese; (2) developed a new model of unknown Chinese words based on characters in the word; (3) increased robustness by adding adaptive modification of pruning thresholds and smoothing of word emission probabilities. While the enhancements to the parser are important for robustness and accuracy, it is even more important to train grammars matched to the conditions of use. For example, parsing a Chinese sentence containing full-width punctuation with a parser trained on half-width punctuation reduces accuracy by over 9% absolute F. In English, parsing accuracy is seriously compromised by training a grammar with punctuation and case to process sentences without them.

We developed grammars for English and Chinese trained specifically for each genre by subsampling from available treebanks (for English, WSJ, BN, Brown, Fisher, and Switchboard; for Chinese, CTB5) and transforming them for a particular genre (e.g., for informal speech, we replaced symbolic expressions with verbal forms and remove punctuation and case) and by utilizing a large amount of genre-matched self-labeled training parses. Given these genre-specific parses, we extracted chunks and POS tags by script. We also trained grammars with a subset of function tags annotated in the treebank that indicate case role information (e.g., SBJ, OBJ, LOC, MNR) in order to produce function tags.

4.2 Individual 5W Systems

The English systems were developed for the monolingual 5W task and not modified to handle MT. They used hand-crafted rules on the output of the latent annotation parser to extract the 5Ws.

English-function used the function tags from the parser to map parser constituents to the 5Ws. First the Who, When, Where and Why were extracted, and then the remaining pieces of the sentence were returned as the What. The goal was to make sure to return a complete What answer and avoid missing the object.

English-LF, on the other hand, used a system developed over a period of eight years (Meyers et al. 2001) to map from the parser's syntactic constituents into logical grammatical relations (GLARF), and then extracted the 5Ws from the logical form. As a back-up, it also extracted GLARF relations from another English-treebank trained parser, the Charniak parser (Charniak 2001). After the parses were both converted to the 5Ws, they were then merged, favoring the system that: recognized the passive, filled more 5W slots or produced shorter 5W slots (providing that the WHAT slot consisted of more than just the verb). A third back-up method extracted 5Ws from part-of-speech tag patterns. Unlike *English-function, English-LF* explicitly tried to extract the shortest What possible, provided there was a verb and a possible object, in order to avoid multiple predicates or other 5W answers.

Chinese-align uses the latent annotation parser (trained for Chinese) to parse the Chinese sentences. A dependency tree converter (Johansson and Nuges 2007) was applied to the constituent-based parse trees to obtain the dependency relations and determine top-level predicates. A set of hand-crafted dependency rules based on observation of Chinese OntoNotes were used to map from the Chinese function tags into Chinese 5Ws. Finally, *Chinese-align* used the alignments of three separate MT systems to translate the 5Ws: a phrase-based system, a hierarchical phrase-based system, and a syntax augmented hierarchical phrase-based system. *Chinese-align* faced a number of problems in using the alignments, including the fact that the best MT did not always have the best alignment. Since the predicate is essential, it tried to detect when verbs were deleted in MT, and back-off to a different MT system. It also used strategies for finding and correcting noisy alignments, and for filtering When/Where answers from Who and What.

4.3 Hybrid System

A merging algorithm was learned based on a development test set. The algorithm selected all 5W's from a single system, rather than trying to merge W's from different systems, since the predicates may vary across systems. For each document genre (described in section 5.4), we ranked the systems by performance on the development data. We also experimented with a variety of features (for instance, does "What" include a verb). The best-performing features were used in combination with the ranked list of priority systems to create a rule-based merger.

4.4 MT Systems

The MT Combination system used by both of the English 5W systems combined up to nine separate MT systems. System weights for combination were optimized together with the language

model score and word penalty for a combination of BLEU and TER (2*(1-BLEU) + TER). Rescoring was applied after system combination using large language models and lexical trigger models. Of the nine systems, six were phrased-based systems (one of these used chunk-level reordering of the Chinese, one used word sense disambiguation, and one used unsupervised Chinese word segmentation), two were hierarchical phrase-based systems, one was a string-to-dependency system, one was syntax-augmented, and one was a combination of two other systems. Bleu scores on the government supplied test set in December 2008 were 35.2 for formal text, 29.2 for informal text, 33.2 for formal speech, and 27.6 for informal speech. More details may be found in (Matusov et al. 2009).

5 Methods

5.1 5W Systems

For the purposes of this evaluation[2], we compared the output of 4 systems: *English-Function*, *English-LF*, *Chinese-align*, and the combined system. Each English system was also run on reference translations of the Chinese sentence. So for each sentence in the evaluation corpus, there were 6 systems that each provided 5Ws.

5.2 5W Answer Annotation

For each 5W output, annotators were presented with the reference translation, the MT version, and the 5W answers. The 5W system names were hidden from the annotators. Annotators had to select "Correct", "Partial" or "Incorrect" for each W. For answers that were Partial or Incorrect, annotators had to further specify the source of the error based on several categories (described in section 6). All three annotators were native English speakers who were not system developers for any of the 5W systems that were being evaluated (to avoid biased grading, or assigning more blame to the MT system). None of the annotators knew Chinese, so all of the judgments were based on the reference translations.

After one round of annotation, we measured inter-annotator agreement on the Correct, Partial, or Incorrect judgment only. The kappa value was 0.42, which was lower than we expected. Another surprise was that the agreement was lower for When, Where and Why (κ=0.31) than for Who or What (κ=0.48). We found that, in cases where a system would get both Who and What wrong, it was often ambiguous how the remaining W's should be graded. Consider the sentence: "He went to the store yesterday and cooked lasagna today." A system might return erroneous Who and What answers, and return Where as "to the store" and When as "today." Since Where and When apply to different predicates, they cannot both be correct. In order to be consistent, if a system returned erroneous Who and What answers, we decided to mark the When, Where and Why answers Incorrect by default. We added clarifications to the guidelines and discussed areas of confusion, and then the annotators reviewed and updated their judgments.

After this round of annotating, κ=0.83 on the Correct, Partial, Incorrect judgments. The remaining disagreements were genuinely ambiguous cases, where a sentence could be interpreted multiple ways, or the MT could be understood in various ways. There was higher agreement on 5W's answers from the reference text compared to MT text, since MT is inherently harder to judge and some annotators were more flexible than others in grading garbled MT.

5.3 5W Error Annotation

In addition to judging the system answers by the task guidelines, annotators were asked to provide reason(s) an answer was wrong by selecting from a list of predefined errors. Annotators were asked to use their best judgment to "assign blame" to the 5W system, the MT, or both. There were six types of system errors and four types of MT errors, and the annotator could select any number of errors. (Errors are described further in section 6.) For instance, if the translation was correct, but the 5W system still failed, the blame would be assigned to the system. If the 5W system picked an incorrectly translated argument (e.g., "baked a moon" instead of "baked a cake"), then the error would be assigned to the MT system. Annotators could also assign blame to both systems, to indicate that they both made mistakes.

Since this annotation task was a 10-way selection, with multiple selections possible, there were some disagreements. However, if categorized broadly into 5W System errors only, MT errors only, and both 5W System and MT errors, then the annotators had a substantial level of agreement (κ=0.75 for error type, on sentences where both annotators indicated an error).

[2] Note that an official evaluation was also performed by DARPA and BAE. This evaluation provides more fine-grained detail on error types and gives results for the different approaches.

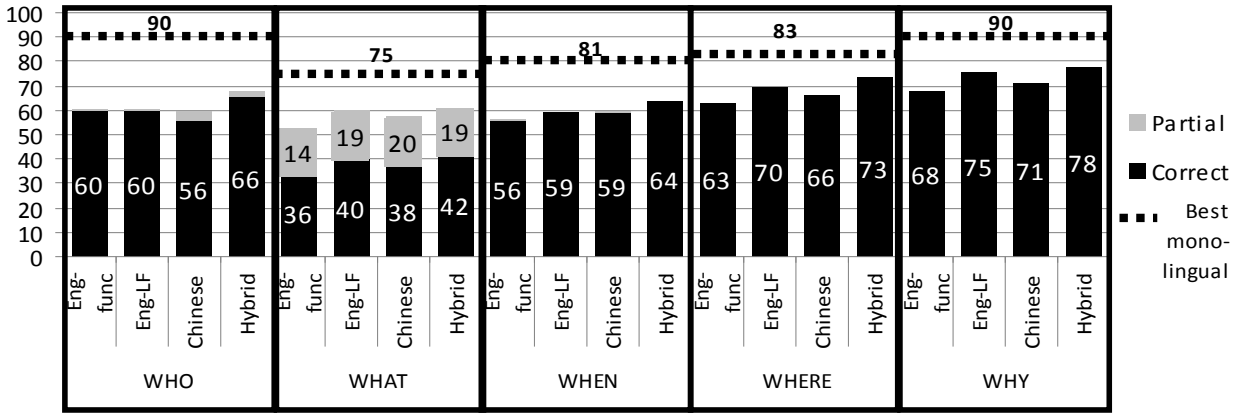

Figure 1. System performance on each 5W. "Partial" indicates that part of the answer was missing. Dashed lines show the performance of the best monolingual system (% Correct on human translations). For the last 3W's, the percent of answers that were Incorrect "by default" were: 30%, 24%, 27% and 22%, respectively, and 8% for the best monolingual system

5.4 5 W Corpus

The full evaluation corpus is 350 documents, roughly evenly divided between four genres: formal text (newswire), informal text (blogs and newsgroups), formal speech (broadcast news) and informal speech (broadcast conversation). For this analysis, we randomly sampled documents to judge from each of the genres. There were 50 documents (249 sentences) that were judged by a single annotator. A subset of that set, with 22 documents and 103 sentences, was judged by two annotators. In comparing the results from one annotator to the results from both annotators, we found substantial agreement. Therefore, we present results from the single annotator so we can do a more in-depth analysis. Since each sentence had 5W's, and there were 6 systems that were compared, there were 7,500 single-annotator judgments over 249 sentences.

6 Results

Figure 1 shows the cross-lingual performance (on MT) of all the systems for each 5W. The best monolingual performance (on human translations) is shown as a dashed line (% Correct only). If a system returned Incorrect answers for Who and What, then the other answers were marked Incorrect (as explained in section 5.2). For the last 3W's, the majority of errors were due to this (details in Figure 1), so our error analysis focuses on the Who and What questions.

6.1 Monolingual 5W Performance

To establish a monolingual baseline, the English 5W system was run on reference (human) translations of the Chinese text. For each partial or incorrect answer, annotators could select one or more of these reasons:

- Wrong predicate or multiple predicates.
- Answer contained another 5W answer.
- Passive handled wrong (WHO/WHAT).
- Answer missed.
- Argument attached to wrong predicate.

Figure 1 shows the performance of the best monolingual system for each 5W as a dashed line. The What question was the hardest, since it requires two pieces of information (the predicate and object). The When, Where and Why questions were easier, since they were null most of the time. (In English OntoNotes 2.0, 38% of sentences have a When, 15% of sentences have a Where, and only 2.6% of sentences have a Why.) The most common monolingual system error on these three questions was a missed answer, accounting for all of the Where errors, all but one Why error and 71% of the When errors. The remaining When errors usually occurred when the system assumed the wrong sense for adverbs (such as "then" or "just").

	Missing	Other 5W	Wrong/Multiple Predicates	Wrong
REF-func	37	29	22	7
REF-LF	54	20	17	13
MT-func	18	18	18	8
MT-LF	26	19	10	11
Chinese	23	17	14	8
Hybrid	13	17	15	12

Table 3. Percentages of Who/What errors attributed to each system error type.

The top half of Table 3 shows the reasons attributed to the Who/What errors for the reference corpus. Since *English-LF* preferred shorter answers, it frequently missed answers or parts of

answers. *English-LF* also had more Partial answers on the What question: 66% Correct and 12% Partial, versus 75% Correct and 1% Partial for *English-function*. On the other hand, *English-function* was more likely to return answers that contained incorrect extra information, such as another 5W or a second predicate.

6.2 Effect of MT on 5W Performance

The cross-lingual 5W task requires that systems return intelligible responses that are semantically equivalent to the source sentence (or, in the case of this evaluation, equivalent to the reference).

As can be seen in Figure 1, MT degrades the performance of the 5W systems significantly, for all question types, and for all systems. Averaged over all questions, the best monolingual system does 19% better than the best cross-lingual system. Surprisingly, even though *English-function* outperformed *English-LF* on the reference data, *English-LF* does consistently better on MT. This is likely due to its use of multiple back-off methods when the parser failed.

6.3 Source-Language vs. Target-Language

The Chinese system did slightly worse than either English system overall, but in the formal text genre, it outperformed both English systems.

Although the accuracies for the Chinese and English systems are similar, the answers vary a lot. Nearly half (48%) of the answers can be answered correctly by both the English system and the Chinese system. But 22% of the time, the English system returned the correct answer when the Chinese system did not. Conversely, 10% of the answers were returned correctly by the Chinese system and not the English systems. The hybrid system described in section 4.2 attempts to exploit these complementary advantages.

After running the hybrid system, 61% of the answers were from *English-LF*, 25% from *English-function*, 7% from *Chinese-align*, and the remaining 7% were from the other Chinese methods (not evaluated here). The hybrid did better than its parent systems on all 5Ws, and the numbers above indicate that further improvement is possible with a better combination strategy.

6.4 Cross-Lingual 5W Error Analysis

For each Partial or Incorrect answer, annotators were asked to select system errors, translation errors, or both. (Further analysis is necessary to distinguish between ASR errors and MT errors.) The translation errors considered were:

- Word/phrase deleted.
- Word/phrase mistranslated.
- Word order mixed up.
- MT unreadable.

Table 4 shows the translation reasons attributed to the Who/What errors. For all systems, the errors were almost evenly divided between system-only, MT-only and both, although the Chinese system had a higher percentage of system-only errors. The hybrid system was able to overcome many system errors (for example, in Table 2, only 13% of the errors are due to missing answers), but still suffered from MT errors.

	Mistrans-lation	Deletion	Word Order	Unreadable
MT-func	34	18	24	18
MT-LF	29	22	21	14
Chinese	32	17	9	13
Hybrid	35	19	27	18

Table 4. Percentages of Who/What errors by each system attributed to each translation error type.

Mistranslation was the biggest translation problem for all the systems. Consider the first example in Figure 3. Both English systems correctly extracted the Who and the When, but for

MT: After several rounds of reminded, I was a little bit Ref: After several hints, it began to come back to me. 经过**几番提醒**,**我回**忆**起来了一点点**。
MT: The Guizhou province, within a certain bank robber, under the watchful eyes of a weak woman, and, with a knife stabbed the woman. Ref: I saw that in a bank in Guizhou Province, robbers seized a vulnerable young woman in front of a group of onlookers and stabbed the woman with a knife. **看到**贵**州省某**银**行内**,**劫匪在众目睽睽之下**,抢夺**一个弱女子**,**并且**,**用刀刺**伤该**女子**。
MT: Woke up after it was discovered that the property is not more than eleven people do not even said that the memory of the receipt of the country into the country. Ref: Well, after waking up, he found everything was completely changed. Apart from having additional eleven grandchildren, even the motherland as he recalled has changed from a socialist country to a capitalist country. **那**么**醒来之后却**发现**物是人非**,**多了十一个**孙子**不**说,连**祖国也从**记忆**当中的社会主**义**国家**变成**了**资本主义国家

Figure 3 Example sentences that presented problems for the 5W systems.

What they returned “was a little bit.” This is the correct predicate for the sentence, but it does not match the meaning of the reference. The Chinese 5W system was able to select a better translation, and instead returned “remember a little bit.”

Garbled word order was chosen for 21-24% of the target-language system Who/What errors, but only 9% of the source-language system Who/What errors. The source-language word order problems tended to be local, within-phrase errors (e.g., “the dispute over frozen funds” was translated as “the freezing of disputes”). The target-language system word order problems were often long-distance problems. For example, the second sentence in Figure 3 has many phrases in common with the reference translation, but the overall sentence makes no sense. The watchful eyes actually belong to a “group of onlookers” (deleted). Ideally, the robber would have “stabbed the woman” “with a knife,” rather than vice versa. Long-distance phrase movement is a common problem in Chinese-English MT, and many MT systems try to handle it (e.g., Wang et al. 2007). By doing analysis in the source language, the Chinese 5W system is often able to avoid this problem – for example, it successfully returned “robbers” “grabbed a weak woman” for the Who/What of this sentence.

Although we expected that the Chinese system would have fewer problems with MT deletion, since it could choose from three different MT versions, MT deletion was a problem for all systems. In looking more closely at the deletions, we noticed that over half of deletions were verbs that were completely missing from the translated sentence. Since MT systems are tuned for word-based overlap measures (such as BLEU), verb deletion is penalized equally as, for example, determiner deletion. Intuitively, a verb deletion destroys the central meaning of a sentence, while a determiner is rarely necessary for comprehension. Other kinds of deletions included noun phrases, pronouns, named entities, negations and longer connecting phrases.

Deletion also affected When and Where. Deleting particles such as “in” and “when” that indicate a location or temporal argument caused the English systems to miss the argument. Word order problems in MT also caused attachment ambiguity in When and Where.

The “unreadable” category was an option of last resort for very difficult MT sentences. The third sentence in Figure 3 is an example where ASR and MT errors compounded to create an unparseable sentence.

7 Conclusions

In our evaluation of various 5W systems, we discovered several characteristics of the task. The What answer was the hardest for all systems, since it is difficult to include enough information to cover the top-level predicate and object, without getting penalized for including too much. The challenge in the When, Where and Why questions is due to sparsity – these responses occur in much fewer sentences than Who and What, so systems most often missed these answers. Since this was a new task, this first evaluation showed clear issues on the language analysis side that can be improved in the future.

The best cross-lingual 5W system was still 19% worse than the best monolingual 5W system, which shows that MT significantly degrades sentence-level understanding. A serious problem in MT for systems was deletion. Chinese constituents that were never translated caused serious problems, even when individual systems had strategies to recover. When the verb was deleted, no top level predicate could be found and then all 5Ws were wrong.

One of our main research questions was whether to extract or translate first. We hypothesized that doing source-language analysis would be more accurate, given the noise in Chinese MT, but the systems performed about the same. This is probably because the English tools (logical form extraction and parser) were more mature and accurate than the Chinese tools.

Although neither source-language nor target-language analysis was able to circumvent problems in MT, each approach had advantages relative to the other, since they did well on different sets of sentences. For example, *Chinese-align* had fewer problems with word order, and most of those were due to local word-order problems.

Since the source-language and target-language systems made different kinds of mistakes, we were able to build a hybrid system that used the relative advantages of each system to outperform all systems. The different types of mistakes made by each system suggest features that can be used to improve the combination system in the future.

Acknowledgments

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In COLING-ACL '98: Proceedings of the Conference, held at the University of Montréal, pages 86–90.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), pages 152–164.

Eugene Charniak. 2001. Immediate-head parsing for language models. In Proceedings of the 39th Annual Meeting on Association For Computational Linguistics (Toulouse, France, July 06 - 11, 2001).

John Chen and Owen Rambow. 2003. Use of deep linguistic features for the recognition and labeling of semantic arguments. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan.

Katrin Erk and Sebastian Pado. 2006. Shalmaneser – a toolchain for shallow semantic parsing. Proceedings of LREC.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. Computational Linguistics, 28(3):245–288.

Daniel Gildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA.

Mary Harper and Zhongqiang Huang. 2009. Chinese Statistical Parsing, chapter to appear.

Aria Haghighi, Kristina Toutanova, and Christopher Manning. 2005. A joint model for semantic role labeling. In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), pages 173–176.

Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In Proceedings of Treebanks and Lexical Theories.

Evgeny Matusov, Gregor Leusch, & Hermann Ney: Learning to combine machine translation systems. In: Cyril Goutte, Nicola Cancedda, Marc Dymetman, & George Foster (eds.) Learning machine translation. (Cambridge, Mass.: The MIT Press, 2009); pp.257-276.

Adam Meyers, Ralph Grishman, Michiko Kosaka and Shubin Zhao. 2001. Covering Treebanks with GLARF. In Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources. Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 51-58.

Teruko Mitamura, Eric Nyberg, Hideki Shima, Tsuneaki Kato, Tatsunori Mori, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, Tetsuya Sakai, Donghong Ji, and Noriko Kando. 2008. Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access. In Proceedings of the Seventh NTCIR Workshop Meeting.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 776–783.

Kristen Parton, Kathleen R. McKeown, James Allan, and Enrique Henestroza. Simultaneous multilingual search for translingual information retrieval. In Proceedings of ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, CA, 2008.

Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007).

Sudo, K., Sekine, S., and Grishman, R. 2004. Cross-lingual information extraction system evaluation. In Proceedings of the 20th international Conference on Computational Linguistics.

Honglin Sun and Daniel Jurafsky. 2004. Shallow Semantic Parsing of Chinese. In Proceedings of NAACL-HLT.

Cynthia A. Thompson, Roger Levy, and Christopher Manning. 2003. A generative model for semantic role labeling. In 14th European Conference on Machine Learning.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In Dekang Lin and Dekai Wu, editors, Proceedings of EMNLP 2004, pages 88–94, Barcelona, Spain, July. Association for Computational Linguistics.

Xue, Nianwen and Martha Palmer. 2005. Automatic semantic role labeling for Chinese verbs. InProceedings of the Nineteenth International Joint Conference on Artificial Intelligence, pages 1160-1165.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese Syntactic Reordering for Statistical Machine Translation. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 737-745.

Jianqiang Wang and Douglas W. Oard, 2006. "Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval," in 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 202-209.