The ISL EDTRL System

Juergen Reichert, Alex Waibel

Interactive Systems Laboratories,
Carnegie Mellon University and Karlsruhe University
juergen@ira.uka.de, ahw@cs.cmu.edu

Abstract

For the translation of text and speech, statistical methods on one side and interlingua based methods on the other have been used successfully. However, the former requires programming grammars for each language, plus the design of an interlingua, while the latter requires the collection of a large parallel corpus for every language pair. To alleviate these problems, we propose an approach that combines the advantages from both worlds. The proposed approach makes use of English or enriched English as an interlingua and can cascade data-driven translation systems into and from this interlingua. We show that enriching English with linguistic information that is automatically derived i only on English data performs better than pure cascaded systems.

1. Introduction

In recent years, a number of translation approaches have been proposed to provide for reliable meaningful translation of text and of speech from one language to another. These include direct approaches (statistical, example-based), transfer and interlingua based approaches. Translation performance is usually the one (if not the most important) consideration for the evaluation of these systems, but has to be balanced by considerations of robustness and portability as additional important dimensions to the translation problem. The translation of speech, in particular, is faced with both of these additional challenges: the input speech and its recognition is fragmentary, ill-formed and errorful, and speech translation systems are frequently required to handle multiple language pairs and language directions to allow for successful cross-lingual dialogs between humans.

To accommodate these additional constraints a popular approach has been the interlingua approach to translation. Here, an intermediate representation of meaning is chosen to express the key idea or intent of the speaker. An input sentence is parsed in terms of its key semantic content, represented in an interlingua structure, and from there an equivalent sentence is generated in another language. The use of an interlingua has several advantages. First, adding a new languages to an existing system is simplified, since a new language has to translate only into and out of the interlingua, and we do not require separate translators for every other language pair the system supports. Second, the translation step extracts only the key intentions from a speaker's utterance, thereby handling colloquial expressions, and reducing the sensitivity to redundancies, and disfluencies in spoken language. Third, the system can generate paraphrases from the interlingua back into one's own language to provide meaningful feedback and verification of the translation, before it is delivered into another language.

The advantages, however, come at a price: The extraction of the key content from a disfluent input sentence requires the development of semantic grammars that extract key information into frames, concepts and slots. Both, the design of a suitable, unambiguous and language independent interlingua as well as the development of grammars that map sentences to meaning are domain-dependent and have to be repeated for each topic or domain. Their development is labor intensive and requires both linguistic expertise and command of the language at hand. As an attempt to solve these problems, automatic learning is proposed to alleviate the manual development work. The most popular approaches at present are statistical and example-based methods. Both extract direct mappings from input to output language using large parallel corpora between these languages. Statistical machine translation permits automatic statistical learning to build a translator rather than manual programming. But a system has to be developed for each language pair. Each translator, in turn, requires a large parallel corpus for training. While parallel corpora are generally available for large common languages, it is rare to find large parallel corpora for more unusual language pairs (say, Paschtu-Catalan) and domains.

In this paper, we therefore develop an alternative strategy: the use of general English and a linguistically enriched English as interlingua. Here we avoid the manual design of an interlingua, and the writing of grammars for analysis and generation; but we also avoid the need for large parallel corpora for every language pair. Moreover, English as interlingua can be 'enriched' by linguistic information extracted in a data-driven fashion automatically and monolingually in English, where plenty of data exists.

2. Description of the EDTRL System

In this section we describe the Error Driven Translation Rule Learning (EDTRL) translation system. The EDTRL system uses enriched English as an interlingua to translate from a source language into a target language going through described special interlingua as an intermediary step. This approach tries to combine the advantages of a system with an explicit interlingua and the advantages of a pure data-driven system. Thereby it becomes possible to add a new language to a given system with n languages very fast by only adding 2 components instead of n-1. The use of enriched English as an interlingua eliminates the need for an explicit, handcrafted interlingua specification and removes the domain limitation which is typical for interlingua-based translation systems.

An additional benefit of this combination is the reduction of the 'Parallel Data Sparseness Problem'. For most non-English language pairs the amount of parallel text corpora is much smaller than the parallel text corpora from each of

these languages paired with English. Using English as an interlingua can therefore increase the amount of available training data.

2.1. Basic Design Ideas of EDTRL

The EDTRL system is based on statistical transfer rules which are automatically learned from bilingual corpora. While the system can learn transfer rules from two non English languages and acts like a direct data-driven translation system, it is designed to use augmented, formalized English as Interlingua. Thus one language of the parallel corpus has to be English, and has to be standardized and annotated with additional linguistic information. The annotation and standardization process only depends on the English part of the parallel corpus and is consequently independent of the source and target language of the system. Annotations made on the English side are projected through the word nad phrase alignment models onto the source and target language. Some mapping errors are introduced by transferring the structural knowledge from English to some other language, but often that can be compensated through the higher quality and quantity of the available structured knowledge in English compared to most other languages.

To allow for the use of a translation system in real-word-applications a small footprint in memory and space as well as a fast translation process are important. To achieve these goals we decided instead to keep the whole statistical translation model to generate statistical rules from the model. Even these generated rules are to many to build a small and fast system, therefore we keep only a subset of all rules generated during the training and use an evaluation test set to determine the most significant and important rules. This allows us to find the best compromise between size and performance for each application domain. The use of probabilistic translation rules makes it easy to add new rules and even exceptions of existing rules. It also allows tracking translation errors and correcting them if necessary. This ability also leads to an interactive learning modus, where the user can teach the system and optimize its behavior.

2.1.1. Standardized and Simplified English

The standardization step tries to map alternative expressions with similar or equal meanings to the most common used alternative. Furthermore the sentence structure is simplified [SE]. E.g. more complex rarely used tenses are replaced by easier ones:

He had spoken. → He spoke.
He would be speaking. → He would speak.

These kinds of simplifications of course remove information, but often such fine nuances are of little value to the quality of the translation given the current state of the MT systems. In most cases the translation profits from the transformations through more reliable alignments and better utilization of the training data.

Even humans can benefit from Simplified English in some technical domains [AECMA]. Sometimes English utterances have some freedom in word order without changing the main meaning of the utterance. To obtain a consistent word order some simple rules are applied:

E.g. please give me … → give me … please

2.1.2. Preserve translation alternatives

The translation errors from the intermediate English to the target language can be reduced if not only the best hypothesis, but additional information from the search is used. We examined the following methods:

n-best list of complete translations: The translation system produces up to *n* alternative translation hypotheses and passes them to the second translation step. The number of hypotheses has to be kept small to guarantee fast overall decoding, thereby allowing only for little variability.

n-best word or phrase alternatives to the best hypothesis: This method selects the single best hypothesis from the first translation step, but augments it by adding alternative words or phrases, which have high translation probabilities.

Full lattice: In order not to fix one translation hypothesis as the basis for constructing these alternatives, we can also pass on full translation lattices. Using a lattice as input for the second translation step has been shown as the most profitable way to use translation alternatives to improve the translation quality.

2.1.3. Additional knowledge sources

Besides translation alternatives, further information on the structure and the semantic content of a sentence can be helpful. Therefore we incorporated the following additional knowledge sources into our system to provide information for the translation process:

Morphological Analyzer: Starting from the WordNet ontology [WordNet] we built a system to analyse an English word form and determine its base form and derivation rule. The analyzer contains a set of common transformation rules and an even larger list of exceptions from these rules. In the current implementation, each word is analysed without using its context or information from former sentences. The precision for finding the base class is over 95% while the determination of the derivation rules is not yet that good.

Sense Guesser: The sense guesser tries to find the sense of a word. Many words have different meanings depending on the context in which they occur. E.g. table can have the senses ‘desk’ or ‘chart’. Often the context of the word can be used for disambiguation. In our example, the context ‘in the’ assigns table to the chart-class, while ‘on the’ assigns it to the desk-class. We used the sense hierarchy from WordNet.

Synonym Generator: WordNet also lists synonyms for words, all within the well structured and linked hierarchy. Both Sense Guesser and Synonym Generator only use open word classes like nouns, verbs, adjectives, and adverbs.

Part-of-Speech Tagger: a statistical Part-of-Speech tagger was used to provide POS-tags. The tagger uses the tag set which is described in [Brill 1995] and was trained on the tagged Brown Corpus.

Named Entity Tagger: a prototype of some handwritten rules allows us to find named entities, which often should be treated in a special way.

Further knowledge sources like sentence type, active or passive voice, politeness, domain or category could also be added.

2.1.4. *Probabilities and Confidence Measures*

Besides the information from the different sources also a probability or confidence measure for each of the knowledge sources and alternatives is added to the Interlingua. Therefore words and phrases carry attributes with possibilities and their possible alternatives. All this combined additional information forms the interlingua (intermediate representation) for the EDTRL system.

For translating into English the interlingua can easily be transformed into plain English by stripping off all additional information and using the most likely alternative. For translating from English into some other language the additional information can be added directly, i.e. transforming plain English into the annotated form which is then used as the interlingua.

2.2. Training and Translating

In the ideal case the learning process only needs parallel texts and optional dictionaries to and from English, because all other knowledge sources operate on English and are independent from the input and output language. However available direct parallel texts or dictionaries from the source to the target language could be incorporated into the system.

2.2.1. *Statistical Alignment*

In a first step a word alignment (IBM1 or modified IBM2) is performed. In a second step a phrase alignment based on the word alignment is executed, which simultaneously joins similar regions on the word alignment matrix and splits the matrix into smaller parts. For these splitting and joining operations normalized probabilities from the word alignment and the language models are used. The phrase alignment generates a collection of partitions of the word alignment matrix and their probabilities.

2.2.2. *Weight Functions for the Alignment*

To enhance the quality of the statistical alignment, weight functions are introduced, which change the weights of a sentence alignment in a special manner according to a heuristic concept. Different weight functions are examined.

A) The *Weight Position Factor* takes into account, that in parallel sentences the source word positions are not independent from the corresponding translation word positions. Often they lie next to the diagonal of the alignment matrix. The following formula can give them a higher weight.

$$\frac{k}{\left|WordAPos - \left(WordBPos \bullet \frac{\#WordsA}{\#WordsB}\right)\right|}$$

B) The *Length Penalty* consider the assumption that longer utterance often results in less accurate alignments and so they are punished using the following expression

$$\frac{k}{\log(len)}$$

C) Parallel utterances of significant different length often produce alignments of minor quality. Therefore the *Matching Length Factor* prefers utterances with almost the same length.

$$k\frac{\#LenA+\#LenB}{2\cdot\max(\#LenA,\#LenB)}$$

D) The *Frequency Weight* keep in mind, that alignments between words with similar frequency are typically more accurate than between words with very different frequency counts.

$$k\frac{\#WordsA+\#WordsB}{2\cdot\max(\#WordsA,\#WordsB)}$$

Each function is parameterized and its parameters are estimated on a validation set. The *Weight Position Factor* gives the far best improvement for the alignment quality, compared to a manually alignment. It reduces the alignment error by 13.1% while the other weight functions give an improvement from 1.5% to 3%. A combination of all four alignment functions reduces the error by 14.6%.
Besides the four weight functions many other functions are imaginable and can be examined.

2.2.3. *Rule Generation and Selection*

On the basis of the phrase alignment, optional dictionaries and the semantic and morphologic knowledge translation rules are generated. Optimal rules should be accurate (not introducing errors in other translation contexts) and should not be too specific, so that they can be applied frequently. Rules are of the form:

Cond1 | *Cond2* | … → *Templ1* | *Templ2* | …

where *Cond* can be a word or phrase containing attribute classes and *Templ* is a template which has to be instantiated during the translation process. Both *Cond* and *Templ* carry probabilities. Most attribute classes are part of a hierarchy. This allows enforcing a match by walking up the tree to a more common representation while at the same time decreasing the rule score. A set of meta-rules controls the construction process. Every time a translation rule contradicts with the training data, the rule is split and new attributes are added to resolve the error.

In order not to get too many rules, each rule is checked for its efficiency on a validation set.

2.2.4. *The Translation Process*

The translation process tries to match and instantiate rules along the input utterance. This results in a search tree which needs to be pruned if it grows too large in size. A beam-search then gets the best hypothesis weighted by a trigram language model. Both directions, to and from the interlingua, are very similar, which is shown in the following simple example. In each direction explicit language knowledge is only used for the English part.

A) Translation: Chinese -> IL (Tagged English):

Input:
我 从 某人 那 传染 上 感冒 了

Rules:
我 从 某人 那 <1> 上 <2> 了 ⇨ I've <VB> a <Disease> from someone 0.6
传 染 ⇨ infection <NN> 0.3 | transmission <NN> 0.1 | infect <VB> 0.2 | catch <VB> 0.1
感冒 ⇨ cold <Disease> 0.3 | rheum <Body Substance> 0.2 | to catch cold <VB,Change> 0.4

Instantiation of the first rule:
=> I think I've caught a cold from someone

B) IL (Tagged English) -> Chinese (or Spanish) :

Input:
I think I've caught <VB> a cold <Disease> from someone

Rules:
I've <VB> a <Disease> from someone
⇨ 我 从 某人 那 <1> 上 <2> 了 0,7
catch <VB> -⇨ 捕捉 0.4, 逮 0.3, 传染 0.1 …
catch <NN> ⇨ 陷阱 0.1, …
cold <Temperature attribute> ⇨ 冷 0.4 , 凉 0.4
cold <Disease> ⇨ 感冒 1.0

Instantiation of the first rule:
=> 我 从 某人 那 捕捉 上 感冒 了

3. Experiments

To evaluate the concept of English as an interlingua we chose Chinese as input language and Spanish as output language, since, in spite of the widespread use of these languages, comparatively few direct Chinese-Spanish translations are available.

We trained a number of translation systems that translate directly from Chinese to English, English to Spanish, and Chinese to Spanish, respectively, and compared the results from the direct Chinese to Spanish systems with two combined approaches that use English as a intermediate language: First, we simply cascaded the Chinese-English and English-Spanish systems, feeding the output of the former into the latter ones. We then translated the same test set using the full EDTRL system's definition of an augmented, formalized version of English as an interlingua. For further comparison, the direct and cascaded translation steps were also done with Systran's publicly available online machine translation system [Systran2004].

The data for these experiments were taken from the Basic Travel Expression Corpus (BTEC), a multilingual collection of conversational phrases in the travel domain [Takezawa2002]. The Chinese-English system was trained on 162316 parallel phrases. As only a subset of 6027 phrases was available in Spanish, only the corresponding parallel phrases were used to train the English-Spanish and Chinese-Spanish systems. The test set consisted of 506 new sentences created for the 2003 CSTAR evaluation campaign, and the scores were calculated using 16 English and in average of 3-4 Spanish reference translations. We report the NIST score using the mteval script [MTeval2002] in version 11.

Systems	EDTRL	Systran
C → E	7.34	5.74
E → S	5.17	6.06
C → S	3.17	-
C → E → S	3.41	2.84
C → E_{IL} → S	3.69	-

Table 1: Results (NIST-Score)

The higher scores of the statistical systems on Chinese to English, compared to translations to Spanish, mainly reflect the facts that a much larger amount of training material was used and that the evaluation was performed with a higher number of references. Surprisingly, the cascaded EDTRL systems resulted in better performance than a directly trained system. This effect is caused by the fact that the EDTRL system uses dictionaries for Chinese-English and English-Spanish, while for Chinese-Spanish no dictionary is available.

Using augmented and formalized English (E_{IL}) as an interlingua in the EDTRL system is shown to yield improvements over the pure cascaded translation.

The results of a slightly improved system for the Chinese-English unrestricted track of the IWSLT 2004 evaluation are given below. The subjective scores are the average of the medians of the three grades assigned to each translation.

Method	Score	Rank (n of 9)
fluency	2.93	6
adequacy	3.25	3
BLEU	0.27	5
GTM	0.66	4
NIST	7.50	2
PER	0.42	3
WER	0.53	4

Table 2: IWSLT 2004 Chinese-English unrestricted

4. Acknowledgement

This work was partially supported by the European Union under the integrated project TC-STAR, IST-2002-2.3.1.6.

5. References

[AECMA] http://www.simplifiedenglish-aecma.org/Simplified_English.htm

[Brill1995] Eric Brill, A Case Study in Part of Speech Tagging, 1995 Association for Computational Linguistics

[MTeval2002] NIST MT evaluation kit version 11. http://www.nist.gov/speech/tests/mt/.

[SE] http://www.userlab.com/SE.html

[Systran2004] http://www.systranbox.com/systran/box

[Takezawa2002] Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. Proc. of LREC 2002, pp. 147–152, Las Palmas, Canary Islands, Spain, May 2002.

[WordNet] Professor George A. Miller, Dr. Christiane Fellbaum, Randee Tengi, Susanne Wolff, Pamela Wakefield, Helen Langone, Benjamin Haskell, WordNet a lexical database for the English language, http://www.cogsci.princeton.edu/~wn/