MATREX: DCU Machine Translation System for IWSLT 2006

Nicolas Stroppa, Andy Way

National Centre for Language Technology
Dublin City University
Dublin 9, Ireland
{nstroppa,away}@computing.dcu.ie

Abstract

In this paper, we give a description of the machine translation system developed at DCU that was used for our first participation in the evaluation campaign of the International Workshop on Spoken Language Translation (2006).

This system combines two types of approaches. First, we use an EBMT approach to collect aligned chunks based on two steps: deterministic chunking of both sides and chunk alignment. We use several chunking and alignment strategies. We also extract SMT-style aligned phrases, and the two types of resources are combined.

We participated in the Open Data Track for the following translation directions: Arabic-English and Italian-English, for which we translated both the single-best ASR hypotheses and the text input. We report the results of the system for the provided evaluation sets.

1. Introduction

In this paper, we present the Data-Driven MT system developed at DCU, MATREX (Machine Translation using Examples). This system is a hybrid system which exploits both EBMT and SMT techniques to extract a dataset of aligned chunks [1].

The EBMT data resources are extracted using a two-step approach. First, the source and the target sentences are chunked using several different methods. In the case of English and Italian, we employ a marker-based chunker [2, 3]. In the case of Arabic, we use the chunker provided with the ASVM toolkit [4]. The chunks are then aligned thanks to a dynamic programming algorithm which is similar to an edit-distance algorithm while allowing for block movements [5, 6]. This aligner relies on the identification of relationships between chunks, which can be defined and computed in several ways. We also extract SMT-style aligned phrases from word alignments, as described in [7]. These two types of resources are then combined and given to the decoding module, currently a wrapper around a phrase-based SMT decoder.

We participated in the Open Data Track for the following translation directions: Arabic-English and Italian-English, for which we translated both the single-best ASR hypotheses and the text input. We report the results of the system for the provided evaluation sets.

This paper is organized as follows. In section 2, we describe the various components of the system; in particular, we give details about the various chunking and chunk alignment strategies. In Section 3, we report experimental results obtained for the two language pairs. In Section 4, we conclude, and provide avenues for further research.

2. The MaTrEx System

The MATREX system used in our experiments is a modular data-driven MT engine, built following established Design Patterns [8]. It consists of a number of extendible and re-implementable modules, the most important of which are:

- *Word Alignment Module*: takes as its input an aligned corpus and outputs a set of word alignments.
- *Chunking Module*: takes in an aligned corpus and produces source and target chunks.
- *Chunk Alignment Module*: takes the source and target chunks and aligns them on a sentence-by-sentence level.
- *Decoder*: searches for a translation using the original aligned corpus and derived chunk and word alignments.

The Word Alignment and the Decoder modules are currently wrappers around existing tools, namely GIZA++ [9] and PHRAMER.[1] In our experiments we investigated a number of different chunking and alignment strategies which we describe in more detail in what follows.

An overview of the entire translation process is given in Figure 1: the aligned source-target sentences are passed in turn to the word alignment, chunking and chunk alignment modules, in order to create our chunk and lexical example databases. These databases are then given to the decoder to translate new sentences.

[1] http://www.utdallas.edu/~mgo031000/phramer/.

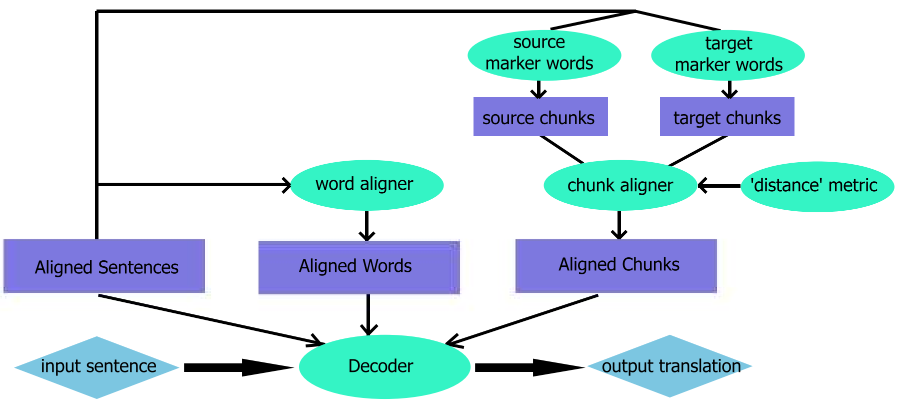

Figure 1: Translation Process

2.1. Chunking Strategies

2.1.1. *Marker-Based Chunking*

One method for the extraction of chunks, used in the creation of the example database, is based on the Marker Hypothesis [3], a psycholinguistic constraint which posits that all languages are marked for surface syntax by a specific closed set of lexemes or morphemes which signify context. Using a set of closed-class (or "marker") words, such as determiners, conjunctions, prepositions, possessive and personal pronouns, aligned source-target sentences are segmented into chunks [2] during a pre-processing step. A chunk is created at each new occurrence of a marker word, with the restriction that each chunk must contain at least one content (or non-marker) word. In addition to the set of marker words used in the experiments of [2, 10], punctuation is also used to segment the aligned sentences – with the punctuation occurring in chunk-final, rather than initial, position. An example of such a chunking is given in Figure 2, for English and Italian.

2.1.2. *Arabic Chunking*

The language characteristics of Arabic makes the direct application of the Marker-Based chunker described above more difficult. In the case of Arabic, determiners, prepositions, and pronouns do not usually form independent tokens but are usually part of a token which also contains a noun, an adjective, or a verb. Consequently, in order to identify the markers, one would need to perform some disambiguation at different levels, in particular tokenization and POS tagging. We would thus lose one of the main strengths of the Marker-Based approach, which is simplicity.

Another option is to use an already available chunker, such as ASVM [4]. This choice is also motivated by our previous work on Basque-English MT [6], in which we used a chunker specifically designed for Basque: we found that the chunks obtained in this manner are actually comparable to the chunks extracted with the marker-based chunker. The ASVM toolkit is based on Support Vector Machines, a Machine Learning algorithm, and has been trained on the Arabic Penn Treebank [11]. The chunking process is achieved through a pipeline approach: tokenization, lemmatisation, POS tagging, and finally chunking are performed in turn.

2.1.3. *Remarks about Chunking*

Since each module of the system can be changed independently of the others, it is possible to use a variety of chunkers. The Marker-Based approach has several obvious advantages: it is simple (linear complexity), easily adaptable, and does not need expensive training on Treebanks. Adapting this chunker to a new language simply amounts to providing the system with a list of marker words. For example, in the case of Italian, we easily extracted a list of markers from the MorphIt lexicon [12], making it possible to apply the Marker-Based chunker to Italian.

However, we do not exclude the possibility to use different types of chunkers that may be available. In particular, in the case of English, several statistical chunkers have been developed, notably in the context of the CoNLL 2000 shared task [13].

2.2. Alignment Strategies

2.2.1. *Word alignment*

Word alignment is performed using the GIZA++ statistical word alignment toolkit and we followed the "refined" method of [7] to extract a set of high-quality word alignments from the original uni-directional alignment sets. These along with the extracted chunk alignments were passed to the translation decoder.

2.2.2. *Chunk alignment*

In order to align the chunks obtained by the chunking procedures described in Section 2.1, we make use of a "edit-distance style" dynamic programming alignment algorithm.

In the following, a denotes an alignment between a target sequence e and a source sequence f, with $I = |e|$ and $J =$

English: *[it felt okay] [after the game] [but then] [it started turning black-and-blue] [is it serious ?]*
Italian: *[era a posto] [dopo la partita] [ma poi] [ha cominciato] [a diventare livida] [è grave ?]*

Italian: *[nel mio caso] [solitamente] [per affari] [raramente] [per piacere]*
English: *[in my case] [it is usually] [on business] [seldom] [for pleasure]*

Figure 2: English and Italian Marker-Based chunking

$|f|$. Given two sequences of chunks, we are looking for the most likely alignment $\hat{a}$:

$$\hat{a} = \operatorname*{argmax}_a \mathbb{P}(a|e,f) = \operatorname*{argmax}_a \mathbb{P}(a,e|f). \qquad (1)$$

We first consider alignments such as those obtained by an edit-distance algorithm, i.e.

$$a = (t_1, s_1)(t_2, s_2)\dots(t_n, s_n),$$

with $\forall k \in \llbracket 1, n \rrbracket$, $t_k \in \llbracket 0, I \rrbracket$ and $s_k \in \llbracket 0, J \rrbracket$, and $\forall k < k'$:

$$\begin{gathered} t_k \leq t_{k'} \text{ or } t_{k'} = 0, \\ s_k \leq s_{k'} \text{ or } s_{k'} = 0, \\ I \subseteq \cup_{k=1}^{n} \{t_k\},\, J \subseteq \cup_{k=1}^{n} \{s_k\}, \end{gathered}$$

where $t_k = 0$ (resp. $s_k = 0$) denotes a non-aligned target (resp. source) chunk.

We then assume the following model:

$$\mathbb{P}(a,e|f) = \Pi_k \mathbb{P}(t_k, s_k, e|f) = \Pi_k \mathbb{P}(e_{t_k}|f_{s_k}), \qquad (2)$$

where $\mathbb{P}(e_0|f_j)$ (resp. $\mathbb{P}(e_i|f_0)$) denotes an "insertion" (resp. "deletion") probability.

Assuming that the parameters $\mathbb{P}(e_{t_k}|f_{s_k})$ are known, the most likely alignment is computed by a simple dynamic-programming algorithm.[2] Moreover, this algorithm can be easily adapted to allow for block movements or "jumps", following the idea introduced in [5] in the context of MT evaluation. This adaptation can be necessary if the order of constituents is significantly different in the source and target languages. In our previous work, we found out that it was useful in the case of Basque-English [6], but not for Spanish-English [1]. In our experiments, we thus decided to include this adapted algorithm for Arabic, but not for Italian.

Instead of using an Expectation-Maximization algorithm to estimate these parameters, as commonly done when performing word alignment [9, 14], we directly compute these parameters by relying on the information contained within the chunks. The conditional probability $P(e_{t_k}|f_{s_k})$ can be computed in several ways. In our experiments, we have considered three main sources of knowledge: (i) word-to-word translation probabilities, (ii) word-to-word cognates and (iii) chunk labels, which are described in the following sections.

2.2.3. *Knowledge Source Combination*

These sources of knowledge can be combined using a log-linear framework, in the following manner:

$$P(e_i|f_j) = \frac{\exp(\sum \lambda_k h_k(e_i, f_j))}{Z}, \qquad (3)$$

where $h_k(e_i, f_j)$ represents a given source of knowledge related to the chunks e_i and f_j, λ_k the associated weight parameter and Z a normalization parameter. The different models are described in more detail below.

2.2.4. *Word-to-word probabilities*

As a criterion to relate chunks, we use word-to-word probabilities, which are simply extracted from the word alignment module, as described above. Relationships between chunks are then computed thanks to the following model, similar to IBM model 1 [14].

$$h_p(e_i, f_j) = \log \prod_k \sum_l \mathbb{P}(e_{i_l}|f_{j_k}). \qquad (4)$$

This model is often used in SMT as a feature of a log-linear model; in this context, it is called a word-based lexicon model [15].

2.2.5. *Cognate identification*

It is also possible to take into account a feature based on the identification of cognates. This is especially useful for texts with technical terms, for which it is possible to identify a significant number of cognates. We use the notation:

$$C(e_{i_l}, f_{j_k}) = \begin{cases} 1 & \text{if there is a cognate between } e_{i_l} \text{ and } f_{j_k}, \\ 0 & \text{otherwise.} \end{cases}$$

We then use the following feature:

$$h_c(e_i, f_j) = \frac{1}{k} \sum_k \max_l C(e_{i_l}, f_{j_k}), \qquad (5)$$

which computes the ratio between the number of cognates identified between the source and the target words, and the total number of source words.

2.2.6. *Chunks label*

If a label is assigned to chunks during the chunking process, we can compare the labels in the source and the target and

[2] This algorithm is actually a classical edit-distance algorithm in which distances are replaced by opposite-log-conditional probabilities.

use this information to relate chunks. In this case, the feature is a simple binary feature:

$$h_l(e_i, f_j) = \begin{cases} 1 & \text{if } e_i \text{ and } f_j \text{ share the same label,} \\ 0 & \text{otherwise.} \end{cases}$$

The weights of the log-linear model are not optimized; we experimented with different sets of parameters and did not find any significant difference as long as the weights stay in the interval $[0.5 - 1.5]$. Outside this interval, the quality of the model decreases.

2.2.7. Integrating SMT data

Whilst EBMT has always made use of both lexical and phrasal information [16], it is only recently that SMT has moved towards the use of phrases in their translation models and decoders [7, 17]. It has, therefore, become harder than ever to identify the differences between these two data-driven approaches [10]. However, despite the convergence of the two paradigms, recent research [10, 18] has shown that by combining elements from EBMT and SMT to create hybrid data-driven systems capable of outperforming the baseline systems from which they are derived. Therefore, SMT phrasal alignments are also added to the aligned chunks extracted by the chunk alignment module, in order to produce higher quality translations.

2.3. Decoder

The decoding module is capable of retrieving already translated sentences and also provides a wrapper around PHRAMER, a phrase-based SMT decoder. This decoder implements Minimum-Error-Rate Training [9] within a log-linear framework [19]. The BLEU metric [20] is optimized using the provided development set. We use a log-linear combination of several common feature functions: phrase translation probabilities (in both directions), word-based translation probabilities (lexicon model, in both directions), a phrase length penalty and a target language model [21].

The phrase translation probabilities are simply estimated thanks to relative frequencies computed on the aligned dataset of chunks obtained as described above. Word-based translation probabilities are introduced to smooth the phrase translation probabilities, that tend to be over-estimated for phrases that appear only once in the training data [15].

The target (English) language model is a simple 3-gram language model trained on the English portion of the training data, using the SRI Language Modeling Toolkit [22], with modified Kneser-Ney smoothing [23].

3. Experimental results

3.1. Data

The experiments were carried out using the provided datasets, extracted from the Basic Travel Expression Corpus (BTEC) [24]. This multilingual speech corpus contains tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad. We participated in the Open Data Track for the following translation directions: Arabic-English and Italian-English, for which we translated both the single-best ASR hypotheses and the text input.

For the supplied data track, 20,000 aligned sentences were provided for training, for both Arabic and Italian. We performed some filtering based on the lengths and the relative lengths of the sentences, ending up with 19,378 aligned sentences for Arabic and 19,599 for Italian. In order to perform MERT, we use the development set number 4, made up of 489 aligned sentences. Note that the system was trained using exclusively the provided datasets.

As a pre-processing step, the English sentences were tokenized using the Maximum-Entropy based tokenizer of the OpenNLP toolkit.[3] This tokenizer was also used for Italian since we found that it was properly dealing with all punctuation marks except apostrophes. For apostrophes, in particular those involved in contractions, we used a set of regular expressions specific to Italian. Additionally, for English and Italian, case information was removed. For Arabic, the tokenization was handled by the ASVM toolkit previously mentioned.

The official metrics of the evaluation campaign of IWSLT 2006 take case information and punctuation marks into account. Since the input sentences do not contain such information, we need to reintroduce them in the output. In order to do so, we followed the procedure suggested by the organizers. For punctuation restoration, we consider that the punctuation marks are hidden events occurring between words, the most likely hidden tag sequence (consistent with the given word sequence) being found using an n-gram language model trained on a punctuated text. For case restoration, the task is viewed as a disambiguation task in which we have to choose between the (case) variants of each word of a sentence. Again, finding the most likely sequence is done using an n-gram language model trained on a case-sensitive text. These 3-gram language models were trained on the English portion of the training data, again using the SRILM toolkit [22].

Since the datasets do not contain named entities, numbers, or acronyms, as an additional post-processing step we removed the words from the output that were copied by the decoder from the input.

3.2. Results

The system output is evaluated with respect to the following metrics: BLEU, NIST, Meteor, WER, and PER. These metrics are computed thanks to the IWSLT 2006 evaluation server. The results are reported in Tables 1 and 2. Official (resp. additional) results includes (resp. excludes) case and

[3] `http://opennlp.sourceforge.net/`.

		BLEU	NIST	Meteor	WER	PER
ASR (1-best)	Official	0.2598	6.585	0.5497	0.5835	0.4869
	Additional	0.2783	7.228	0.5495	0.5662	0.4498
Text input	Official	0.3126	7.546	0.6246	0.5315	0.4286
	Additional	0.3467	8.358	0.6245	0.4964	0.3744

Table 1: Official results - Italian

		BLEU	NIST	Meteor	WER	PER
ASR (1-best)	Official	0.145	4.531	0.402	0.7027	0.5949
	Additional	0.1391	4.794	0.4	0.7165	0.5870
Text input	Official	0.1624	4.89	0.4336	0.686	0.5678
	Additional	0.1589	5.29	0.432	0.6935	0.5537

Table 2: Official results - Arabic

punctuation information.

The results obtained show that our system is competitive with other start-of-the-art systems, which is encouraging for our first participation. Moreover, we successfully adapted our system to Italian-English, a new language pair. As expected, the results obtained on the text input are better than those obtained on the ASR (1-best) output, for almost all of the metrics, the difference ranging from 1.7 (Arabic, Official) to 6.8 (Italian, Additional) BLEU points. With respect to Official vs. Additional, we excepted to get better results on the Additional metrics, since it seemed to be an easier task. This is what we obtained for Italian: 3.41 additional BLEU points on the text input. However, it is less clear for Arabic, for which the Additional scores were better, except for NIST, which conflicts with the other metrics.

4. Conclusion

In this paper, we described MaTrEx, the hybrid Data-Driven MT system developed at DCU. This system was used for our first participation in the evaluation campaign of IWSLT 2006. This system uses both EBMT and SMT approaches to extract aligned chunk resources. We described several chunking and chunk alignment strategies, integrated within a modular system.

We participated in the Open Data Track for the Arabic to English and Italian to English translation tasks, for which we translated both the single-best ASR hypotheses and the text input. We showed that our system can be easily adapted to new language pairs, and is competitive with other state-of-the-art systems.

We plan to continue our experiments in various directions. First, we will investigate different language pairs, in particular Chinese-English. Then, we will examine how to combine different chunking strategies. We also want to explore the use of other chunk alignment techniques.

5. Acknowledgments

This work is supported by Science Foundation Ireland (grant number OS/IN/1732).

6. References

[1] S. Armstrong, M. Flanagan, Y. Graham, D. Groves, B. Mellebeek, S. Morrissey, N. Stroppa, and A. Way, “Matrex: Machine translation using examples,” in *TC-STAR OpenLab on Speech Translation*, Trento, Italy, 2006.

[2] N. Gough and A. Way, “Robust large-scale EBMT with marker-based segmentation,” in *Proceedings of TMI 2004*, Baltimore, Maryland, 2004, pp. 95–104.

[3] T. Green, “The necessity of syntax markers. two experiments with artificial languages,” *Journal of Verbal Learning and Behavior*, vol. 18, pp. 481–496, 1979.

[4] M. Diab, K. Hacioglu, and D. Jurafsky, “Automatic tagging of arabic text: From raw text to base phrase chunks,” in *Proceedings of HLT-NAACL 2004*, Boston, MA, 2004, pp. 149–152.

[5] G. Leusch, N. Ueffing, and H. Ney, “CDER: Efficient MT evaluation using block movements,” in *Proceedings of EACL 2006*, Trento, Italy, 2006, pp. 241–248.

[6] N. Stroppa, D. Groves, A. Way, and K. Sarasola, “Example-based machine translation of the Basque language,” in *Proceedings of AMTA 2006*, Cambridge, Massachusetts, 2006, pp. 232–241.

[7] P. Koehn, F. Och, and D. Marcu, “Statistical phrase-based translation,” in *Proceedings of HLT-NAACL 2003*, Edmonton, Canada, 2003, pp. 48–54.

[8] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1995.

[9] F. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of ACL 2003*, Sapporo, Japan, 2003, pp. 160–167.

[10] D. Groves and A. Way, "Hybrid example-based SMT: the best of both worlds?" in *Proceedings of the Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, ACL 2005*, Ann Arbor, Michigan, 2005, pp. 183–190.

[11] M. Maamouri and A. Bies, "Developing an arabic treebank: Methods, guidelines, procedures, and tools," in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004*, Geneva, Switzerland, 2004.

[12] E. Zanchetta and M. Baroni, "Morph-it!: a free corpus-based morphological resource for the italian language," in *Proceedings of the Corpus Linguistics 2005 conference*, Birmingham, UK, 2005.

[13] E. F. T. K. Sang and S. Buchholz, "Introduction to the conll-2000 shared task: Chunking," in *Proceedings of CoNLL 2000 and LLL 2000*, Lisbon, Portugal, 2000, pp. 127–132.

[14] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[15] R. Zens and H. Ney, "Improvements in phrase-based statistical machine translation," in *Proceedings of HLT-NAACL 2004*, Boston, MA, 2004, pp. 257–264.

[16] M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle," in *Artificial and Human Intelligence*, A. Elithorn and R. Banerji, Eds. Amsterdam, The Netherlands: North-Holland, 1984, pp. 173–180.

[17] P. Koehn, "Pharaoh: A beam search decoder for phrase-based statistical machine translation models," in *Proceedings of AMTA 2004*, Washington, District of Columbia, 2004, pp. 115–124.

[18] D. Groves and A. Way, "Hybrid data-driven models of MT," *Machine Translation, Special Issue on EBMT*, 2006, (to appear).

[19] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proceedings of ACL 2002*, Philadelphia, PA, 2002, pp. 295–302.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of ACL 2002*, Philadelphia, PA, 2002, pp. 311–318.

[21] R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney, "The RWTH phrase-based statistical machine translation system," in *Proceedings of IWSLT 2005*, Pittsburgh, PA, 2005, pp. 155–162.

[22] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, 2002, pp. 901–904.

[23] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Harvard University, Tech. Rep. TR-10-98, 1998.

[24] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in *Proceedings of LREC 2002*, Las Palmas, Spain, 2002, pp. 147–152.