Using Monolingual Corpora for Statistical Machine Translation: The METIS System

Yannis Dologlou
Institute for Language and Speech Processing, Artemidos 6 & Epidavrou, Paradissos Amaroussiou 151 25, Athens, Greece.
ydol@ilsp.gr

Stella Markantonatou
Institute for Language And Speech Processing, Artemidos 6 & Epidavrou, Paradissos Amaroussiou 151 25, Athens, Greece.
marks@ilsp.gr

George Tambouratzis
Institute for Language and Speech Processing, Artemidos 6 & Epidavrou, Paradissos Amaroussiou, 151 25, Athens, Greece.
Giorg_t@ilsp.gr

Olga Yannoutsou
Institute for Language and Speech Processing, Artemidos 6 & Epidavrou, Paradissos Amaroussiou, 151 25, Athens, Greece.
olga@ilsp.gr

Athanassia Fourla
Institute for Language and Speech Processing, Artemidos 6 & Epidavrou, Paradissos Amaroussiou, 151 25, Athens, Greece.
soula@ilsp.gr

Nikos Ioannou
Institute for Language and Speech Processing, Artemidos 6 & Epidavrou, Paradissos Amaroussiou, 151 25, Athens, Greece.
nioannou@ilsp.gr

Abstract

In this article the principles of the METIS Machine Translation system are presented. METIS employs an extensive tagged and lemmatised corpus of texts in the target language, coupled with bilingual lexica covering the desired pairs of source-target languages. To generate a high-quality translation, the METIS system is provided with statistical tools enabling it to extract linguistic knowledge from the annotated corpus of the target language. The advantage of this approach is that, although no grammars need to be provided explicitly, grammatical translations are retrieved from the corpus by using pattern matching techniques.

1 Introduction

METIS is an innovative approach to Statistical Machine Translation (SMT) in that it relies on monolingual corpora only (as far as we can ascertain, all known SMT systems rely on bitexts). METIS, at the moment, has the power of a translation memory but there still remains a vast field of research regarding the potential abilities of the system. At the moment we experiment with general corpora.

In Section 3 we present the system: its principles (Section 3.1), the linguistic resources required (Section 3.2) and the algorithms (Section 3.3). In Section 4 we present the first evaluation results. Finally, in Section 5, we present our conclusions and discuss future research. First, however, we start by briefly sketching out the state of the art in SMT (Section 2).

2 A Brief Review of the related Literature

A good percentage of the experimental and almost all commercial Machine Translation (MT) systems rely on linguistic techniques (for this term and relevant review see [Dorr et al., 1999]). Linguistic-based MT typically employs large bilingual computational lexica and grammars of various designs, posing considerable demands on the thoroughness

of the syntactic and semantic information encoded. These resources are mostly handcrafted by specialized linguists. Therefore it could be claimed that linguistic-based MT does not really rely on corpora (however, a system of this type can be fed with rules and lexica automatically extracted from corpora). Significant linguistic-based MT systems and projects include EUROTRA [Johnson et al., 1985], SYSTRAN [Fourla et al., 2000], LOGOS [Scott, 1989].

In recent years, the family of corpus-based MT systems, all of which rely on large, typically parallel and aligned, corpora (or bitexts), has attracted the attention of researchers. The aim in this case is to make use of sentence availability in both the source and target languages (hereafter denoted as SL and TL respectively). Example-based MT [Nagao, 1984] and Statistical MT (SMT) belong to this paradigm.

In their pioneering SMT work, Brown et al. (1990, 1993) rely on the assumptions that (i) a SMT can be built without recourse to any handcrafted (or other) linguistic resource such as bilingual lexica and grammars (ii) thus, a word-to-word model suffices provided that large, well-aligned corpora are used to bootstrap it. Nowadays, there exists a remarkable variety of "aligners" of bitexts at sentence level (for a comprehensive review see [Melamed, 2001]). Brown and his colleagues had not proposed any decoder (a decoder takes a previously unseen SL sentence, s1, and tries to locate the TL sentence, t1, that maximizes the probability P(t1|s1)). Such systems may reach a translation quality similar to existing linguistic-based commercial systems. However, a considerable expertise in mathematics is required for the system design, development and fine-tuning, making these systems inaccessible to a wider audience of scientists (which probably is the main reason why SMT systems have remained unpopular within the wider scientific community).

Recently, modifications of the original idea by Brown et al. (1993) have been proposed, attacking several fronts. On the linguistic resources front, the use of bilingual dictionaries and lists of cognates [Al-Onaizan et al., 1999] and the exploitation of syntactic knowledge such as treebanks [Yamada & Knight, 2001] have been combined with the core idea of aligned bilingual corpora, yielding promising results. Of course, the original argument has been weakened in this way. For instance, if syntactic knowledge is exploited, then the word-to-word assumption is violated.

Another line of improvement concerns the front of alignment and decoding. Aligned bitexts are heavily used for bootstrapping and have been an active research field ([Melamed, 2001]), in an effort to improve translation quality by optimising the alignment of bitexts. However, the fact remains that bitexts are sparse, therefore methods are sought that survive on small bitexts [Al-Onaizan et al, 2000]. The introduction of decoders [Germann et al, 2001] constitutes another line of research, aimed to provide a higher-quality translation, while limiting the time needed to determine this solution. To achieve that, the A* search algorithm is used to locate the optimal solution.

The system proposed here differs from previous SMT work in that it relies solely on monolingual corpora (though monolingual corpora of a target language plus hand-crafted bilingual lexica have been used in the past for word-sense disambiguation [Dagan et al., 1994]), thus offering a radical solution to the problem of sparse bitexts. The system makes use of handcrafted bilingual lexica (including cognates), which add grammatical information in the form of PoS tags. Furthermore, the system uses a set of hand-crafted tag-mapping rules. The intention is to provide the statistical component with a structure that is as close as possible to the target language. This strategy has been shown to be effective in rule-based systems [Dyvik, 1995].

3 METIS principles

3.1 Aim - General Characteristics

Most Statistical Machine Translation (SMT) approaches are intended to discover translations of input text by exploiting very large parallel corpora (also termed as 'bitexts') aligned at some level of granularity, which most often is the sentence level. A statistical model is trained on the bitext and then it is used to translate new sentences.

Evidently, for these approaches to work, it is essential that huge amounts of bitext be available. Such corpora are rare even for the most widely spoken languages, for instance for the pair English/French. At the same time, more and more monolingual corpora of reasonable size are becom-

ing available for an ever-increasing set of languages.

The novelty of the proposed MT system concerns the elimination of the use of bitexts altogether. Instead, the proposed system relies on large monolingual texts while requiring some standard linguistic technology. The translation is achieved by employing:

(i) a set of language-specific resources such as taggers and lemmatisers for both the source and the target language.

(ii) a set of bilingual resources; these are bilingual lexica tuned to the requirements of the particular system and rules for mapping between the tagsets of the source and the target language.

(iii) pattern recognition-inspired statistical models that extract information from large corpora in the target language as opposed to the language models employed by other SMT systems.

Thus, the amount of explicit linguistic information, which is employed to map sentences of the source language on sentences of the target language, is substantially reduced compared to rule-based MT systems. The monolingual corpus serves as a rich pool of information and is processed with a set of 'intelligent' learning rules to extract the correct translations.

Pattern recognition-inspired statistical models have been used quite extensively in written text processing. Applications range from systems for identifying morphological relations [Pentheroudakis et al., 1995] to the automatic generation of thesauri from raw text [Grefenstette, 1993] and the alignment of bitexts [Melamed, 2001].

Learning algorithms allow the proposed system to acquire information from the textual data itself concerning the morphosyntactic properties of the translations and help to resolve any lexical ambiguities. These algorithms would most likely be classified into the statistical pattern recognition and/or artificial intelligence domains. A major issue is the definition of a distance between linguistic material in the source and target languages. Then, initial correspondence decisions are made at a local level, while keeping a set of alternative solutions with their likelihood scores. These local solutions are combined to reach the final solution at a global level, by maximising the likelihood of the global solution. In this way, the rule-based part of machine translation involving syntactic analysis and generation is largely avoided, while only well-established linguistic technology is employed.

3.2 Linguistic Resources

Three types of linguistic resources have been used: corpora, bilingual lexica and sets of structure-adjusting rules. We provide details regarding each one of them in turn.

BNC has been used as a target language corpus. As a source language corpus, we have used the ILSP corpus [Gavrilidou et al., 1998]. Both corpora have been tagged and lemmatized. BNC has been tagged/lemmatized with the MBLEM tool [Van den Bosch and Daelemans, 1999]. The ILSP corpus has been tagged/lemmatized with the ILSP suite of tools [Papageorgiou et al., 2000]. For the BNC, the CLAWS5 tagset has been used, while the ILSP/PAROLE tagset has been used [Lambropoulou et al., 1996] for the ILSP corpus.

A bilingual Greek-English lexicon of approximately 10,000 Greek lemmata has been adapted to the needs of METIS. For each Greek lemma, a number of English translations are supplied, up to a limit of ten. Related expressions are also supplied. The lexicon provides PoS information for both the source and the target language covering single- and multi-word entries. Figure 1 explains the structure of the lexicon and provides an example for a single-word Greek entry that receives more than one English translation as well as an example of an expression.

Thus, the lexica provide both the lexical links between the source and the target language as well as some elementary morphosyntactic information in the form of PoS tags.

Furthermore, we have developed and are currently experimenting with several sets of structure-adjusting rules henceforth referred to as "grammars". These rules handle tags, lemmata and distances between tags during the translation process. The tags convey PoS information as well as some grammatical information such as tense, number, voice, degree etc. A table with the correspondences between the ILSP-Parole and the CLAWS5 tags has been constructed, an extract concerning the nominal paradigm being given in Figure 2.

The structure-adjusting rules are intended:

1. To 'adjust' the source language sentence structure by mapping it onto some structure which is closer to the corresponding

target language one. This step has been shown to greatly enhance the translation quality in rule-based systems [Dyvik, 1995] and seems to fit a pattern-matching based system like the one presented here.

2. To account for multiple structural correspondences. For instance, the present tense of Modern Greek corresponds to both the simple and continuous present tense of English. However, the present tense of Greek is realized with one word while the continuous present tense of English is realized with two words (1).

(1) Το παιδί παίζει.
'The child plays.'
'The child is playing.'

The rules take as input the tags and lemmata corresponding to a given source sentence in the order defined by the sentence and map them onto strings which may or may not differ in the number of tags and/or lemmata to those expected to occur after the application of the bilingual lexica. The left-hand side of the rules constrains the rule application in two ways: (i) by requiring that specific tags/lemmata exist and (ii) by requiring that a certain distance (in words) exists between two specified tags/lemmata. Below, we give the rule that maps the Greek present tense to the English simple and continuous present tenses (2). The rule requires that a verbal present tense exist in the source language string and maps it on two alternative structures, on the simple present tense and on the continuous present tense one. The reentrances here indicate translation equivalence.

(2) [[1]/VbMnIdPr…] -> [[1]/VV?] or [be/VB?] + [[1]/VVG]

3.3 Algorithmic description of the translation process from Language A to Language B.

The basic input unit for the translation process within the METIS system is the period that has been grammatically annotated and lemmatized. Thus, the input of the source language A is a string Aw consisting of ordered lemmata and corresponding grammatical annotations, Aw_i, i=1, …n.

At first, a table lookup process replaces each lemma (or appropriately defined groups of lemmata) of language A with the appropriate lemma(ta) of language B, by means of a bilingual dictionary. In addition, the tags of the source language are mapped on the tags of the target language by means of the same bilingual dictionary. The result is a set of lemmata and tags Bw_i, i = 1, …, m in language B. Furthermore, structure-adjusting rules are applied on the Aw string and result in deletions and/or additions on the Bw string.

Next, the correct permutation of Bw_is can be established by selecting the appropriate sentence structure within a large monolingual corpus of language B. For that purpose the sequence Bw_i , i=1,…m is compared against each one of the phrases of the large monolingual corpus and a distance is computed that takes into account all possible permutations of words in Bw. The phrase of the corpus that provides the smallest distance is selected as the correct translation of Aw. Note that this algorithm will not only provide the best permutation of Bw_i , i=1,…m but also determine necessary additions/deletions of words that are imposed by the structure of the monolingual corpus. The implementation of this step requires the definition of a distance between different words/tags and also the use of a fast algorithm to compute the minimum permutative distance between two sets of words.

4 Innovation of the METIS approach

The METIS approach possesses several advantages over existing MT systems and MT/SMT systems under development:

1. Unlike existing SMT systems, it does not rely on bitexts, which are rare even for the widely spoken languages. Instead, it makes use of monolingual corpora, which are freely/readily available for many languages (and are in any case easier to create).
2. Unlike many other existing SMT systems, METIS does not construct explicitly a statistical language model which it later exploits to retrieve translations. Instead, it uses pattern-matching techniques which determine the similarity between attested structures of the target language and appropriately transformed structures of the source language.
3. Compared to rule-based MT systems, it requires a substantially smaller

amount of language-specific resources and tools: it uses only POS taggers and lemmatizers. These tools exist for several languages and the technology for creating them is well established. Certain basic linguistic resources need to be defined for each language pair: (i) bilingual lexica, which already exist for many language pairs, and (ii) a limited set of rules that define a mapping relation between the tagsets of the source and the target language.

4. Only minimal effort is required to add a new language pair (lexica and tag-mapping rules).

5. METIS allows the fine-tuning of the linguistic model to suit the specific task. An improved performance can be achieved by changing the corpus modules only, while retaining other modules such as the tag-mapping rules and linguistic tools.

5 Evaluating the METIS output

For evaluating the METIS output the following assumptions were made:

- the retrieved sentence may have a structure similar to the target one but the meaning should not be altered,
- if a similar sentence in terms of meaning does not exist, the system will retrieve the most similar one in terms of structure so that certain post-processing rules are applied, and
- optimum quality of the other tools (taggers, lemmatizers and lexica) involved is assumed.

The need for an objective evaluation led us to test the METIS output against human translation, with a view of also comparing it with a translation memory in the near future. For the purposes of the evaluation exercise, a metric system was developed focusing on the quality of the output translations. This metric system provides a range of "penalties" accounting for the significance of errors; for instance a wrong verbal lemma is considered a more serious mistake than a wrong article. The test material comprised: test sentences, target-language corpora, bilingual lexica, tag-mapping rules and adjustment of weights.

Test sentences covered a wide range of linguistic phenomena such as valency, personal-impersonal verb phrases, word order, tense and aspect, subordinate clauses, and specific structural differences between Greek and English. All the toy examples were extracted from the ILSP corpus. The sentences had a low level of complexity and a fixed limit regarding length (up to 8 words).

For example, one toy target corpus consists of 28 sentences: the exact translations of the source sentences (more than one exact translation existing for some source sentences (3)) and sentences which had a varying number of common elements with the exact translations (4).

(3)Source sentence:
Η γυναίκα καθαρίζει το μήλο.
Exact translations:
The woman peels the apple.
The woman is peeling the apple.
The woman cleans the apple.
The woman has been peeling the apple.

(4) Other sentences:
The woman peeled the apple.
The woman peels the pear.
etc..

The sentences in the target corpus were classified according to the translation quality metrics developed for this purpose. Then the system was calibrated to minimize the difference between the order of obtained translations and the original qualitative classification of the target corpus (to reflect that the system behavior should be similar to that of the human translator).

In the example mentioned above the toy lexicon had 136 entries, expressions and punctuation marks included, while the tag mapping rules (5) covered the English translation equivalents of the Greek present tense.

```
(5) <RuleSet\
    [1\\VbMnIdPr_____IpAv__] ->
            [1\\VVB] | [1\\VVZ]|
            [\be\VBB]+[1\\VVG]          |
            [\be\VBZ]+[1\\VVG] |
            [\have\VHB]+[\be\VBN]+
               [1\\VVG]
            [\have\VHZ]+[\be\VBN]+
               [1\\VVG];
```

\RuleSet>

This first evaluation has shown that the system won't miss a good or a near translation, which is present in the corpus, provided that it is not embedded into a much longer period. Furthermore, it was shown that tag-mapping rules are not always necessary: for instance, they certainly improved the system's behaviour as far as verbal tenses are concerned, but it did not have any impact on the handling of clitics and possessives. Of course, due to the METIS principles, all produced translations are grammatically correct.

6 Conclusions - Discussion

In this article, the principles of the METIS Machine Translation system have been presented. In contrast to other machine translation systems, METIS employs only monolingual corpora (in the target language) and bilingual lexica, while it relies on pattern matching techniques. By design, METIS is intended to extract linguistic knowledge from the large annotated corpus of the target language. Thus no detailed grammars are required for analysis, transfer and synthesis. The system is currently being developed for two language pairs, Greek-to-English and Dutch-to-English. Preliminary results with prototypes of the system have illustrated its potential in generating high-quality translations.

Several promising directions for future research have been determined: (i) introducing more structure-adjusting rules and/or fine-tuning the weights in the assignment problem in order to optimise translation accuracy (ii) enhancing the ability of the system to extract bits of existing periods, which better match the input one (iii) fine-tuning the system resources (mainly the corpus and bilingual lexicon) towards a specified sub-language. All these factors need to be further studied and thoroughly evaluated in order to investigate the possibility that METIS system may successfully bridge the gap between translation memories and rule-based systems.

7 Acknowledgement.

This work is partially supported by the Future and Emerging Technologies (FET) unit of the IST programme of the European Commission. The author is solely responsible for the content of this communication. It does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of data appearing therein.

References

BNC: http://www.hcu.ox.ac.uk/BNC/

CLAWS5 tagset: http://www.comp.lancs.ac.uk/ucrel/claws5tags.html

ILSP corpus: http://www.ilsp.gr/hnc_gr.html

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N.A., Yarowsky, D. 1999. Statistical Machine Translation, Final Report, John Hopkins University.

Al-Onaizan, Y., U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, K. Yamada. 2000. Translating with Scarce Resources. In Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence, July 30-August 3, Austin, Texas, pp. 672-678.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer and P. S. Roosin. 1990. A Statistical Approach to Machine Translation. Computational Linguistics, Vol. 16, No. 2, pp 79-85.

Brown, P., S. Della Pietra, V. Della Pietra and R. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". Computational Linguistics, Vol. 19, No. 2, pp 263-312.

Dagan, I. and A. Itai. 1994. "Word Sense Disambiguation Using a Second Language Monolingual Corpus". Computational Linguistics, Vol. 20, No. 4, pp 563-596

Dorr, B.J., Jordan, P.W. and Benoit, J.W. 1999. A Survey of Current Paradigms in Machine Translation. Advances in Computers, Vol. 49, pp. 1-68. London: Academic Press

Dyvik, H., 1995. Exploiting Structural Similarities in Machine Translation. Computers and the Humanities, Vol. 28, pp. 225-234.

Fourla A., Yannoutsou O., Tsakou I., Stamou S. and Petrits A., 2000. The contribution of a user group to the evaluation and improvement of an MT system. Translating and the Computer, Vol. 22. London: ASLIB.

Gavrilidou, M., P. Lambropoulou, N. Papakostopoulou, S. Spiliotopoulou & N. Nassos. 1998 Greek Corpus

Documentation, Parole LE2-4017/10369, WP2.9-WP-ATH-1.

Germann, U., M. Jahr, K. Knight, D. Marcu and K. Yamada. 2001. Fast Decoding and Optimal Decoding for Machine Translation. In Proceedings of the Conference of the Association for Computational Linguistics (ACL 2001), Toulouse, France, pp. 228-235.

Grefenstette, G. (1995) Comparing two Language Identification Schemes. In Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95), Rome, Italy, Dec. 1995. Available at http://www.rxrc.xerox/com/publis/mltt/jadt/jadt.html.

Johnson, R., M. King and L. des Tombe. 1985. EUROTRA: a Multilingual System under Development. Computational Linguistics, Vol. 11, April-September 1985, pp. 155-169

Labropoulou, P., E. Mantzari, and M. Gavrilidou. 1996. Lexicon-Morphosyntactic Specifications: Language Specific Instantiation (Greek), PP-PAROLE, MLAP 63-386 report

Melamed, D.I. 2001. Empirical Methods for Exploiting Parallel Texts. The MIT Press

Nagao, M., 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In A. Elithorn and R. Banerji (eds) Artificial and Human Intelligence. North-Holland

Papageorgiou H., P. Prokopidis, V. Giouli, and S. Piperidis. 2000. "A Unified PoS Tagging Architecture and its Application to Greek". In Proceedings of The Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May-2 June, pp. 1455-1462.

Pentheroudakis, J. and Vanderwende, L. 1993. Automatically Identifying Morphological Relations in Machine-Readable Dictionaries. Technical Report MSR-TR-93-06, Microsoft Research Advanced Technology Division, Microsoft Corporation, One Microsoft Way, Redmond, WA. 98052.

Scott, B.E. 1989. The LOGOS System. In Proceedings of MT SUMMIT II, pp. 174-179. Deutsche Gesellschaft fuer Dokumentation, e.V. (DGD).

Van den Bosch, A. and W. Daelemans. 1999. Memory-based morphological analysis. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99, University of Maryland, USA, June 20-26, 1999, pp. 285-292.

Yamada, K. and K. Knight. 2001. A syntax-based statistical translation model. Proceedings of the Conference of the Association for Computational Linguistics.

Source language word/phrase	Source language tags	English word/phrase	English tags
(source language string) lemmata for the declinable parts, otherwise the word form, with hashes between words in the case of phrases	POS tags In the case of phrases, tags are supplied for the declinable parts only, with asterisks otherwise, while hashes are used to indicate spaces between words.	(English string) lemmata for the declinable parts, otherwise the word form with hashes between words in the case of phrases	CLAWS5/BNC In the case of phrases, tags are supplied for the declinable parts only, with asterisks otherwise, while hashes are used to indicate spaces between words.
εκτελώ	VbPv	execute	VV?
εκτελώ	VbPv	be#accomplished	VB?#*
εκτελεστικό#απόσπασμα	Aj#No	firing#squad	*#NN?

Figure 1. Format of the bilingual lexica

PoS	Type of Feature	ILSP-Parole	CLAWS5
GR: No **EN : N**	Noun type :common, Proper	Cm, Pr	N, P0
	Gender (Masculine,Feminine,Neutral)	Ma,Fe,Ne	
	Number (Unspecified number, Singular, Plural)	Sg,Pl	Nosymbol, 1,2, for common nouns only
	Case (Nominative, Genitive ,Dative, Accusative, Vocative)	Nm,Ge,Da,Ac,Vo	

Figure 2. Correspondences between ILSP/Parole and CLAWS5 tagsets for the nominal paradigm