Recent Advances in Intelligent Information Systems
ISBN 978-83-60434-59-8, pages 751–760

Genetic Algorithm-based Multi-Word Automatic Language Translation

Ali Zogheib

IT-Universitetet i Goteborg - Department of Applied Information Technology

Abstract

An Automatic Language Translation System's quality depends mainly on that of two components: the *Alignment approach* and the *Translation Model*. In this paper, we will present an alignment approach that covers *one-to-one*, *one-to-many*, *many-to-one*, *many-to-many alignment*, whose output is used by a translation model based on *Genetic Algorithm*. The *Translation Model* searches for the decomposition of a phrase, into *single-* and *multi-word units*, that gives the best translation, while allowing for units' containment and overlapping. The system was used on 17475 *English-French* phrases from the European Parliament' debates.

Keywords: Automatic Language Translation, Machine Learning, Genetic Algorithm, Alignment Metric, single- and multi-word linguistic units Alignment.

1 Introduction

Automatically translating a sentence $S\{w_i\}$, from a source language to a target one, can be described as the process of replacing each source word w_i, with its corresponding word in the target language. Or, in practice, some words need to be dropped, some to be added and others need their translations' orders, in the target language, to be reversed. Many methods were proposed dealing with these exceptions, the most famous ones are the IBM models. We propose a method that solves these exceptions, by dealing with *multi-word* units' identification and alignment. With this method, the sentence S becomes an ordered set $\{U_k\}$ of *single/multi-word* units that may overlap. Or, there are many ways to decompose a sentence into units; an automatic translator should select the decomposition that gives the best possible translation. In our Automatic Translation System, we used the *Genetic Algorithm* (*GA*) (Holand, 1992) to find the best translation. We believe that our system is the first that use of *GA* in the translation process.

In the next section, we will present the *multi-word* units approach, the motivations behind it, what type of units to identify and how to align them. In section 3, the proposed *Translation Model* is described in terms of *GA's* structure and search. Experiments' results will be presented and commented in section 4.

2 Multi-Word Units

Each pair of languages imposes a set of constraints on the automatic language translation system, but all these pairs share a subset of constraints that can be grouped into the following three context-related constraints:

- The constraint of *ordering the translated words*: For the *English-French* pair, the order of some translated words may be reversed in the other language, for example: *United Nations/Nations Unies*.
- The second constraint is *the NULL generated translations*. They are the words that are needed to be present in the translation, without being the translation of any other words, like *de* and *la* in the pair: *Commission Proposal/Proposition de la Commission*.
- The reverse of the previous constraint. *Some words need to be dropped* while translating to a target language (i.e. in the reverse of the previous pair, *de* and *la* must disappear), if we are translating from *French* to *English*.

Many approaches (Brown et al., 1993, Cherry et al., 2003, Yamada et al., ...) were proposed, that dealt with these constraints. Some incorporated words' presence, absence and order in the target language, directly in the alignment and translation processes (Cherry et al., 2003, Brown et al., 1993, Och et al., 2000, ...), others managed these constraints by hierarchical alignment (Watanabe et al., 2002, Yamada et al., 2001, ...), or by clustering words (Moore, 2005), Others dealt indirectly with these constraints using statistical models for phrases identification (Koehn 2003). Our approach is designed to satisfy these contextual constraints, by searching and aligning contiguous set of words, that we call *multi-word units*[1], from the corpus and this for both, source and target languages, without incorporating any information on words' translation order, presence/absence in the target language. In the following sections, we will present our approach's motivations and the related algorithms.

2.1 Multi-Word Motivations

Ideally, an Automatic Translation System is a system that models the knowledge used by human translators, and uses it to generate translations, automatically. Taking this objective as our system's goal, we proceeded with an analysis of how skilled human translators do their task. Four critical points were observed:

When faced with a new sentence, (1) they do not start to produce the translation on word by word basis. To the contrarily, (2) they seek to identify linguistic units, formed of contiguous words (*multi-word units*), if they find such ones, (3) they invoke their memory to identify for these found units their corresponding translations, (4) without making any analysis of their constituents' (words) order, absence or presence in each know unit's translation. They already handled these units before, many times; They are frequent units.

[1] In this paper, we used *multi-word* unit terminology to emphasize the fact that we are targeting a sub set of what is refereed to as *phrase* in the literature; more exactly, those satisfying the above mentioned constrains.

In our *Alignment Model*, we modeled the human translator's ability to know that a contiguous set of words in a sentence constitutes a significant unit, by a *Multi-word Unit*'s *Identification Algorithm*. The algorithm scans the corpus for both language, to identify the possible units and that preserves those appeared frequently in each part of the corpus. The knowledge of each identified unit's translation(s), in the target language, is modeled by a *Multi-word Units Alignment Algorithm*. For each found frequent unit in the source language, the algorithm aligns the probable translation(s), from the frequent units found in the target language. Each of these two algorithms will be presented in the following sections.

2.2 Units Identification Algorithm

Human translators' acquisition process of common *multi-word* units in each language, in particular the critical points (2) and (4), was modeled by an algorithm that simulates the acquisition process, as much as possible. No hypotheses are made on the units' structures. It is based on the idea that a corpus does always contain a set of frequent multi-word units, generally domain-dependent or commonly used units in the language.

The algorithm for *Multi-word units* is based on the textitsingle-word one. This later consists, simply on scanning separately for each language, the corpus for the different words it contains. For *multi-word units*, the algorithm follows these steps:

1. Scan the corpus, for the current language, for the distinct words it contains.
2. Let L be the unit's length[2]
3. For L varying from 2 to L_{max}
4. For each phrase in the corpus:
   - Extract each L consecutive words, from the phrase
   - Increase the appearance count of the identified units by 1

For *The European Parliament adopted the amendements* sentence, and $L = 2$, the algorithm will generate the following units: *The European, European Parliament, Parliament adopted, adopted the, The amendements.*

The algorithm can be optimized or replaced by other more sophisticated algorithms. In the current implementation, our aim was to have an unsupervised mean to extract/identify frequent units from the corpus, for each language, without regard to the offline time this identification may require.

2.3 Units Alignment Algorithm

The identified frequent multi-word units, from the corpus, for a source language, are supposed to have their translations frequent in the second half of the corpus, and thus be within the identified multi-word units in the target language.

The alignment of *single-word* units, in the source language, to their possible single-word units' translations, in the target language, can be done using any alignment algorithm. The *multi-word Alignment Algorithm*, we are proposing uses the single-word units' alignment results, to align the multi-word units.

[2]number of contiguous words it contains.

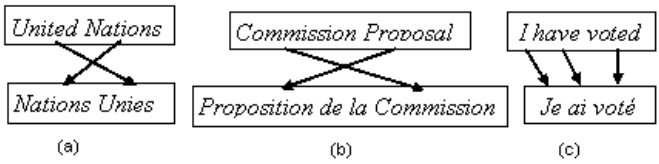


FIGURE 1: Multi-Word Units Alignment

2.3.1 Motivations

The subset of multi-word units that are of interest for us, those that we targeted, are the units whose translations are composed by the translations of each word in the source unit, if any, without regard to their orders. Our idea is direct and simple, and consists of defining, for each language the set of words that may be added/dropped, while translating to/from this language, when seen within a multi-word unit. We call these words *Connectivity Words*. For *French*, we included {*de, des, du, le, la, les, l'*} and for *English* {*of, the, to*}.

With this definition, we can align correctly phrases like the ones in Figure 1, based on units statistics and word alignment results; and that, without including any information on word's order or any complex management of *NULL* generated words for most of cases, at least for *English-French* pair.

2.3.2 Generic Multi-word Units Alignment

After aligning the distinct words, of the source language, to their corresponding possible translation, in the target language, using a alignment algorithm of choice; the *multi-word* units *Generic Alignment Algorithm*, consists of the following steps:

For each identified source unit, U_S, to align:

1. Remove the connectivity words from the unit, if any
2. Identify the target units $\{U_T\}$, who co-occurred with U_S, in the corpus.
3. Remove connectivity words from each unit in $\{U_T\}$, if any
4. Filter-out units from $\{U_T\}$ whose remaining words were not aligned[3] to any of the remaining words in unit U_S, with an alignment metric[4] value above a chosen threshold $Th_{Align-Metric}$.
5. Apply a *one-to-one* alignment algorithm of choice, on the pairs $\{U_S, U_T\}$.

It is clear that the quality of the alignment, for *multi-word* units, will depend solely on that of the used *one-to-one* alignment algorithm. If the unit's words were aligned wrongly, so will be the alignment of their containing units.

3 Translator Model

The proposed alignment model, is an unsupervised algorithm, and the only control we have on its output, is to keep the alignments whose metric values were above a

[3] By the Alignment Algorithm of choice, on the base of *one-to-one* (*word-to-word*)
[4] we used the alignment metric proposed in (Zogheib, 2007)

threshold. Doing so, we filter out un-realistic alignments, without ensuring that the remaining alignments are all correct neither we can guarantee that the alternatives, target units aligned to the same source unit, can be equally treated. There is a need to a model for constructing the translated sentence, based on multi-word units. These latters bring three problems: A word, in a source sentence to translate, (1) may appear in many source units who themselves (2) may be contained in neighboring units in the sentence or (3) be overlapping with these neighboring units. In this work, we propose a translation model that searches for each sentence, to translate, the best and most accurate translation, and that by finding the decomposition, into single/multi-word units, that gives the sought translation. We explored the possibilities of an Artificial Intelligence search method, suited for this task, the Genetic Algorithm (Holand, 1992).

In what follows, we will present the overall behavior, and how the translation is implemented with GA.

3.1 Behavior

Given a phrase to translate, the translator module tries to translate it as follow:

1. For each word w_s in the phrase, identify all possible units U_S, present in the phrase, such that w_s is in U_S.
2. For each linguistic unit U_S, consult the Units Alignment's results for its all possible aligned units $\{U_T\}$[5].
3. Run the Translator's GA, to find the best combination of the aligned units

3.2 Algorithm

The algorithm, we used to generate for a sentence its best possible translation, the Genetic Algorithm, is one of famous search methods (Holand, 1992). It is based on the principle: the best survive. From randomly generated guesses/translations[6], the algorithm tries to find the best translation, by following the steps:

1. Randomly generate a population of N chromosomes (possible solutions),
2. Evaluate the degree of acceptance (fitness function) of each chromosomes
3. Select two chromosomes, and apply a random exchange of their properties values. We call the newly generated chromosomes, offsprings.
4. Randomly alter offsprings
5. Insert the new offsprings in the next generation
6. Repeat till generating N offsprings
7. Repeat step 2) till the fitness function reaches its maximum value, or has stabilized, or the number of generations has reached a maximum.

[5] Above a specified threshold

[6] We use interchangeably guess, chromosome and translation.

Table 1: Chromosome Structure

$Gene_i$		...	$Gene_N$	
U_{Sp}	U_{Tq}	...	U_{Sp}	U_{Tq}

3.3 Chromosome

The chromosome structure has to deal with many difficulties, relative to *multi-word* linguistic units:

- A word w_s, in a phrase, may appear in many possible units U_S. Thus, there are many possible ways to decompose a sentence into units. The chromosome structure has to allow for dynamic phrase decomposition.
- The phrase decomposition, into units, may not be possible without overlapping. The chromosome has to deal with units overlapping as well as units' containment (units completely inside other units).

To satisfy these two needs, two types of information have to be encoded:

- The decomposition, of the phrase to translate, into linguistic units
- The correspondence between source units $\{U_S\}$ and their aligned units $\{U_T\}$.

Chromosome Coding: In order to respond to the above mentioned requirements, the chromosome is defined as a structure composed of n genes each of which corresponds to a word in the source phrase. Each gene $gene_i$, corresponding to word w_i, at position i in the sentence, contains:

- The index of a known source unit U_{sp}, appearing in the source sentence, where the word w_i, appears within. This index varies from 1 to P, where P is the number of units, within the source phrase, in which the word w_i occurs.
- The index of a unit U_{Tq} aligned with the source unit U_{Sp}. The index varies from 1 to Q, where Q is the number of aligned units to U_{Sp}.

Chromosome Decoding: Converting the chromosome, to a readable translation, is done as follows:

- For each gene $gene_i$ the source and target linguistic units corresponding to U_{Sp} and U_{Sq} are identified,
- If two successive words' $[w_i,w_{i+1}]$ associated units U_{Sr} and U_{St} are, such that:
  - One contains completely the other. In this case, we consider the cooccurring unit U_{Tq} associated to the containing unit.
  - If they do [not] overlap, the corresponding units are preserved.

3.4 Fitness Function

After identifying the target units U_{Tq}, we compute the chromosome's fitness as the product of the alignment metric's values for each pair (U_{Sp},U_{Tq}).

$$fitness = \sqrt[N_u]{\prod_p M(U_{sp}, U_{tq})}$$

English	French
european citizenship	Citoyenneté européenne
european central bank	banque centrale européenne
cultural heritage	patrimoine culturel
fundamental problem	problème essential
luxembourg summit	sommet de luxembourg
national currency	monnaie nationale

(a)

English	*French*
for your speech	pour votre intervention
Many thanks	merci beaucoup
not only	pas seulement
wishes to speak	souhaite intervenir
of transparency	de la transparence
Like to congratulate the rapporteur	voudrais féliciter le rapporteur

(b)

FIGURE 2: Units Identification Alg. and Units Alignment Alg. Results

The *sqrt* order is N_U, the number of units that decomposed the phrase, in the chromosome. Two supplementary constraints were applied:

- If two consecutive source units overlap, their aligned units must also overlap. If they do not do so, the fitness is penalized: *fitnesss*= -1 + *fitness*.
- If two consecutive units do not overlap, then their aligned units must not overlap. If they do not do so, the fitness is penalized: *fitness*= -1 + *fitness*.

4 Experiments

For experiments, we used a subset of the European Parliament (*English-French*) corpus as the training set (17475 phrases in each language). We restricted the sentences to those of medium length (between 7 to 10 words) and in order to filter out given wrongly translated sentences, we imposed a constraint on the relative length difference between each sentence and its given translation (2 or less words, who are expected to be auxiliary words i.e. *do*, *ne*, *pas*, .).

In the following sections, the experiments' results for *Units Identification*, *Units Alignment* and *Sentence Translation* algorithms will be presented.

4.1 Units Quality

Our main aim of identifying the multi-word units, from the corpus, is to identify units whose words translation's order is altered between the source and target languages and/or having Connectivity Words, that should be added or removed in the generated translation. From *Table (a) in Figure 2*, we can observe that the algorithm identified the targeted units. Or, not all identified units were significant. Two types of units were identified by the algorithm: *linguistically meaningful* units (*Tables of Figure 2*), and *insignificant ones* (i.e. *us see*, *call vote*, *well let*, *their early*, *least this*, *terms this*, *iii b*, *this report we*, ...). This later type is to be expected, see that the algorithm is an unsupervised algorithm and identifies all units appearing in the corpus, even those that appeared once. Meaningful units are supposed to appear more frequently in the corpus than the meaningless units. Using a threshold on the unit's frequency, allows to filter out the later ones.

From the *English* text, 38524 multi-word units were identified and 41364 ones from the *French* text. Nearly half of the identified units (English 45.3%, French 44.4%) appeared only two times in the corpus, indicating that the corpus is reach in multi-word units, but with low statistic.

Table 2: Alignment Results for Multi-word Units with cooccurrence count of 10

	En-Fr
Recall	97.8%
Precision	97.8%
AER	2.2%

4.2 Units Alignment

From the 38524 English multi-word units, the alignment algorithm aligned 13928 English units. Having nearly half of the identified units, for each language, appearing twice in the text, we put a restriction on the alignment of un-frequent units, consisting of aligning them, only where possible, to French units having the same spelling. For the frequent units, we used a threshold of 0.04, on the alignment metric (for more details see Zogheib, 2007), that allowed to reduce the number of wrong alignment. Tables (a) and (b) in Figure 2 show examples of the produced alignments.

Evaluating the produced alignments' quality, for 13928 aligned units, is difficult. We took, as a representative sample of the alignment's results, the units who cooccurred with their counterparts, in the French language, 10 times, which constitutes a reasonable statistic for quality evaluation. *Recall*, *Precision* and *Alignment Error Rate* (Och et al., 2003), were the metrics we used to evaluate the Alignment's Quality (Table 2). We can observe that the multi-word units' alignment algorithm correctly aligned units with a high accuracy (above 97%).

Aligning all identified units, is the ideal target for any alignment algorithm. In the current version, the proposed algorithm succeeded to align 30% of the units. Many units were not aligned, cause of their low frequencies.

4.3 Translation

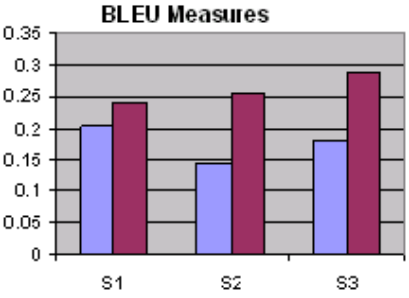
We run the Translation module's GA, with the parameters specified in *Table (a) of Figure 3*, over three Test Sets, each of 50 sentences selected randomly. *Table (c)* presents examples of generated translations, with comparison with those provided in the corpus.

From *Table (c)*, we can observe that each generated translation had fully translated the meaning of the reference sentence, with a very good structure quality, and some grammatical errors. Grammar related issues, weren't targeted in the current implementation. Our aim was to design a system that, when translating a sentence from one language to another, translates correctly the information itself not necessarily the grammar, even if it is always better to have such correct grammar.

Another observation is that, although the generated translation was correct on the meaning and structure levels, it didn't use the same words as the reference sentence: (1) Some words were aligned to the correct word of the opposite gender, *male/female*, (i.e. *ce* vs. *cette*, *le* vs. *la*, ...) or *multiplicity*, *single/plural*, (*va* vs. *vont*, *le* vs. *les*, ...) or in other cases the wrong tense, *past/present*, (*presente* vs. *presenté*). (2) A sentence can be translated using many alternatives for each

Population Size	100
Generations	30
Mutation Prob.	0.10
CrossOverProb	0.75
CrossOver Points	2
Selection Method	Tournament
Tournament Prob.	0.75
Tournament Size	4

(a)



(b)

Phrase	clearly that is necessary for a single market
Given translation	cela est clairement nécessaire en vue du marché unique
Multi-word trans.	clairement ce est nécessaire pour une marché unique
Phrase	that also forms part of the legal framework
Given translation	cela fait aussi partie du cadre juridique
Multi-word trans.	ce également formes partie du cadre juridique
Phrase	that is a matter for the intergovernmental conference
Given translation	cette question relève de la conférence intergouvernementale
Multi-word trans.	ce est une question pour la conférence intergouvernementale

(c)

FIGURE 3: (a) GA's parameters values. (b) Examples of Generated Translations. (c) BLEU measures over the three Test Sets (dark bars for *multi-word* units)

word/unit it contains. Many given translations and their corresponding generated ones did not use the same alternatives, naturally they will not contain the same words. Thus, using an automatic evaluation mean, based on the comparison between the generated translation and the reference sentences' co-occurring words, as do the automatic evaluation metrics, will give mean metric values, not reflecting the real translation's quality; as it can be seen in *Figure 3-b*. *Figure 3-b*, presents the widely adopted metric measured values, the BLEU (Papineni et al. 2002) metric, for all the three Test Sets (S_1, S_2 and S_3). We can observe that for multi-word units, the metric values were promising (from 0.24 to 0.287). They can be enhanced by a factor of at least 2, if the system took the *gender*, *multiplicity* and *tenses*, into account in the translation system. Also, *Table (b) in Figure 3* shows the highest performance of the translation with *multi-word* unit (*dark bars* over *single-word* units, which is to be expected, see that the correct words' order and *NULL* generated words are implicitly embedded in the corresponding aligned units.

5 Conclusions

In this paper, we addressed the problem of *NULL* generated words, and words' order alteration, between source and target languages. We proposed algorithms: For identifying multi-word linguistic units, expected to contain the addressed points, For aligning them to their peers in the target language, and finally For auto-

matically generating translation, using an Artificial Intelligence method, with the Genetic Algorithm. We presented ideas, for future work, that allow taking advantages of these algorithms, particularly the integration of an Automatic Rewriter, as a preprocessing phase and as a post-translation one.

References

Colin Cherry and Dekang Lin: *A Probability Model to Improve Word Alignment*, Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 88-95.

Franz Joseph Och, Hermann Ney: *Statistical Machine Translation.* EAMT Workshop, pp. 39-46, Ljubljana, Slovenia, May 2000.

Franz Joseph Och, Hermann Ney: *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics 2003, 29(1):19-51.

John H. Holland: *Adaptation in Natural and Artificial Systems*, MIT Press, 1992, ISBN: 0262581116.

Kenji Yamada and Kevin Knight: *A Syntax-based Statistical Translation Model.* Meeting of the ACL 2001

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu: *BLEU: a Method for Automatic Evaluation of Machine Translation.* Proc. of the 40th Annual Meeting of ACL, Philadelphia, July 2002, pp. 311-318.

Peter E Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer: *The mathematics of statistical translation: Parameter Estimation.* Computational Linguistics Volume 19, Number 2, 1993

Philipp Koehn, Franz Josef Och, Daniel Marcu: *Statistical Phrase-Based Translation*, HLT/NAACL 2003, p: 48-54.

Robert C. Moore: *Association-Based Bilingual Word Alignment*, Proc. of ACL Workshop on Building and Using Parallel Texts, p: 1-8, Ann Arbor, June 2005.

Taro Watanabe, Kenji Imamura, Eiichiro sumita: *Statistical Machine Translation based on Hierarchical Phrase Alignment*, 9th Int. Conf. on Theoretical and Methodological Issues in Machine Translation, 188-197, 2002.

Zogheib Ali: *Automatic Language Translation - Statistic-based System.* REPORT NO. 2007:4. Chalmers University of Technology, Sweden.