THE INTERNAL RULES OF THE EXAMPLE DATABASE DESIGN

Nadia Luiza DINCA
Research Institute for Artificial Intelligence, Bucharest

Abstract: One of the internal rules of the example-based translation is the dependence of the translation quality on the length and representation of translation examples. These are managed by an example database, for which the linguist should answer two questions when s/he intends to design it: (1) what is the representation chosen for the translation example? and (2) what are the possibilities to generalize the translation examples stored in the database? In this paper my intent is to propose two answers, in fact – a synergetical research direction for Romanian, English and French language. I will consider the representation of the translation examples as depedency trees and I will generalize them by means of the semantic information described by Levin's verb classes.

Keywords: machine translation, translation example, dependency tree, example database design

1. Introduction

The main process of the example-based machine translation is divided into three phases. First, find the most similar examples to the input sentence. Then, recombine the translation of the input sentence according to the most similar example and bilingual dictionary. Lastly, produce the translation of the input sentence.

The example-based machine translation was defined as a translation by analogy which was using an unannotated example data base, created, usually, from a bilingual dictionary - Nagao [1984: 173-180]. The equivalents were represented as word pairs, except the verb equivalents, formalised as case frames.

Later, the structural translation conceived the representation of translation examples as dependency trees with explicited links established between sub-trees (including the leaf nodes, corresponding to the lexical units). These links allow the use of parts of the translation example or sub-trees in order to recognise, for the source language, the exact match between input segments and structures, and for the target language, to select and to combine the equivalent translation units.

MBT2 is the second prototype system in S. Sato and M. Nagao's Memory-based Translation Project. The two researchers introduced the representation called *matching expression*, which represents "the combination of fragments of translation examples. The translation process consists of three steps: (i) make the source matching expression from the source sentence. (ii) transfer the source matching expression into the target matching expression. (iii) construct the target sentence from the target matching expression."

The concept *matching expression* considers three basic operations applied on dependency sub-trees which are already in database: delete the identifier of a certain sub-tree; replace the identifier with a matching expression; add a matching expression as a child of root node of the identifier.

This mechanism generates some candidates of translation. To select the best translation out of them, “a score of a translation was defined, so that it should reflect the correctness of the translation unit. The last is a fragment of a source (or target) word-dependency tree, and also a fragment of a translation example. The more similar these two environments are, the better” (S. Sato and M. Nagao, 1992: 247-252).

The system proposed by H. Kaji et al. in 1992 is a two-phase example-based machine translation methodology which develops translation templates from examples and then translates using template matching.

A translation template is a bilingual pair of sentences in which corresponding units (words and phrases) are coupled and replaced with variables. Conditions concerning syntactic categories, semantic categories, etc. are attached to each variable. A word or phrase satisfying the conditions can be substituted for a variable. The two pseudo-sentences constituting a template include the same set of variables.

The learning procedure is divided into two steps. “In a first step, a series of translation templates is generated from each pair of sentences in the corpus. The first step is subdivided into coupling of corresponding units (words and phrases) and generation of translation templates. In the second step, translation templates are refined to resolve conflicts among them” (Kaji, H. et al.,1992: 672-678).

Translation based on templates consists of (i) source language template matching, (ii) translation of words and phrases and (iii) target language sentence generation. “First, a translation template is retrieved. Words and phrases in the source language sentence are then bound to each variable in the template. Second, the words and phrases which are bound to variables are translated by a conventional machine translation method. Finally, a target language sentence is generated by substituting the translated words and phrases for the variables in the target language part of the translation template” (Kaji, H. et al., ibidem).

In this article, the design of an example database for Romanian, English and French language is realised by following two steps:

- The translation example is represented by the means of dependency trees between which corresponding links are established. At the same time, the types of syntactic dependency relations between the composant units of a verb phrase are identified.
- In order to be generalized, the verb requires a semantic class, by considering Levin’s typology [Levin, 1993]. A gap between the input string and the sub-strings from the example database is filled up by calling the semantic verb class and, implicitly, the verb list which established the synonymous relation between its verbs and the input verb.

2. Representation of the translation examples

2.1. Preliminary discussion

The translation example is a phrase, sometimes having a different meaning not deductible from those of the individual words, and to whom a translation and an exact meaning are assigned for the target language.

A translation example is composed of three parts:

- a source dependency tree (Romanian and English, in this paper);
- a target dependency tree (English and French, in this paper);
- correpondence links.

These three parts are shown in the following verb phrase, extracted from G. Orwell's novel, "1984", subject of a very extended linguistic project, Multext-East:

îşi imaginase orice ↔ had imagined everything ↔ avait tout imaginé
ro_e ([ro1,
[ro1.1., [imagina, v],
[ro1.2, [îşi, pron]],
[ro2, [orice, pron]]]])
en_e ([en1, [had, aux]],
[en2, [imagine, v],
[en3, [everything, pron]]]])
fr_e ([fr1, [avait, aux]],
[fr2, [imaginer, v],
[fr3, [tout, pron]]]])
clinks ([[ro1, en2], [ro2, en3]], [[fr1, en1], [fr2, en2], [fr3, en3]]).

Each number with prefix 'ro', 'en' or 'fr' in the word-dependency trees represents the ID of the sub-tree. Each node in a tree contains a word (in root form) and its syntactic category. A correspondence link is represented as a pair of IDs: *clinks ([[fr1, en1], [fr2, en2], [fr3, en3]])*. A word-dependency (sub)tree which has a correspondence link is translatable; e.g.: el, e2, e3, fr l, fr2, fr3. A translatable tree in which some translatable sub-trees are removed is also translatable; e.g.: e l - e 2 , e 2 - e 3 , e l - e 2 - e 3 , fr l - fr 2, fr2 - fr3, fr1 - fr2- fr3 .

The translation process consists of three steps: decomposition, transfer, and composition [S. Sato and M. Nagao, 1990: 247-252]. In decomposition, the system decomposes a source word-dependency tree into translation units, and makes a source matching expression. In the transfer step, the system replaces every ID in the source matching expression with its corresponding ID. In the composition step, the system composes the target word-dependency tree according to the target matching expression.

2.2. Syntactic Dependency Relations

All the units which constitute the utterance are arranged by the speaker in well-specified constructions, by taking into consideration the dependencies created between them: one word form depends on another for its linear position and its grammatical form [I. Mel'cuk, 2003].

The surface syntactic structure represents a tree whose nodes are labeled with all the lexemes of the sentence, and the arcs receive the names of a language specific syntactic relation, as it is exemplified in the followings lines.

The three major classes of syntactic dependencies, namely: complementation, modification, coordonation, are responsible for a largeg number of syntactic relations at the verb phrase level:

I. Subordinate Syntactic Relations:

a. direct object:

(cumpărase – **ob-dir** → **cartea**) ↔ (bought – **ob-dir** → [the] **book**)
*(bought – **ob-dir** → [the] **book**) ↔ ([avait] acheté – **ob-dir** → [le] **livre**)*
(luă – **ob-dir** → [o] **ţigară**) ↔ (took – **ob-dir** → [a] **cigarette**)
*(took – **ob-dir** → [a] **cigarette**) ↔ (prit – **ob-dir** → [une] **cigarette**)*

b. indirect object in Dative

([să] spună – **ob-indir** → **i**) ↔ ([should] tell – **ob-indir** → **him**)
*([should] tell – **ob-indir** → **him**) ↔ (**lui** ← **ob-indir** – indiquerait)*

c. prepositional object in Accusative

([se simţea] atras– **ob-prep** → **de** [el]) ↔ ([felt] drawn– **ob-prep** → **to** [him])

d. infinitive object

([le] putea – **ob-inf** → **vedea**) ↔ (could– **ob-inf** → **see**)
*(could– **ob-inf** → **see**) ↔ (pouvait – **ob-inf** → [les] **voir**)*

II. Coordinate Syntactic Relation

scoase – **ob-dir**→ [un] **toc** –**coord**→ [o] **sticlă** [de cernează] –**coord**→ **şi**[un volum] ↔ took down **–** **ob-dir**→ [a] **penholder** –**coord**→ [a] **bottle** [of ink] – **coord**→ **and** [a book]
*took down **–** **ob-dir**→ [a] **penholder** –**coord**→ [a] **bottle** [of ink] –**coord**→ **and** [a book] ↔ (sortit [du tiroir] – **ob-dir**→ [un] **porte-plume** –**coord**→ [un] **flacon** [d'encre] –**coord**→ [un] **in-quarto**).*

2.3. Dependency Trees and Correspondence links

Three sets of criteria are used to establish the syntactic relations between two verb phrases, for Romanian, English and French language:

- criteria for syntactic connectedness of two word forms;
- criteria for the syntactic dominance between two word forms;
- criteria for the specific type of the given syntactic dependency between two word forms.

These criteria are language specific relations, which sometimes make the identification of a correspondence difficult. It is the case of the French adverbial pronoun *en*, for example, not realised in English utterance:

îşi turnă o ceaşcă de ceai ↔ pour out a teacupful ↔ en versa une pleine tasse
ro_e ([ro1, [turna, v],
[ro2, ob-indir, [îşi, pron]],
[ro3, ob-dir,
[3.1, [ceaşcă, n],
[ro3.2, [o, art]],
[ro3.3, [de, prep],
[ro3.4, [ceai, n]]]]]])
en_e ([en1,
[en1.1., [pour, v],
[en1.2, jonctiv, [out, prep]],
[en2, ob-dir,
[en2.1, [teacupful, n],
[en2.2, [a, art]]]]]])
fr_e([fr1, [verser, v],
[fr2, ob-adverbial, [en, pron]],
[fr3, ob-dir,

[fr3.1, [tasse, n],
[fr3.2, [une, art]],
[fr3.3, [pleine, adj]]]]])
clinks ([[ro1,en1], [ro3, en2]], [[en1, fr1], [en 2-3, fr2-3]])

The type of syntactic dependency specific to the indirect object in Romanian is undertaken in English by the subject, as the main agent. The type of syntactic dependency proper to the adverbial object is accomplished in English by the semantics of the entire utterance *a teacupful*, while, in French, the pronoun requires knowledge from the previous sentence: *Le liquide répandait une odeur huileuse, écœurante comme celle de l'eau-de-vie de riz des Chinois*. The second source node is extended into daughters 2.1.- *teacupful* and 2.2.- *a*, and it has correspondence links to the third target, but, for a complete understanding, someone should take into consideration the second target node too.

In the following situation, the personal pronoun *her* is not realised in a French equivalent, because its meaning is included in the compositional meaning of the construction *elle tendit*- "subject-predicate":

extended her arms towards the screen ↔ tendit les bras vers l'écran
en_e ([en1, [extend, v],
[en2, dir-obj,
[en2.1., possession, [her, pron]],
[en2.2., [arms, n]]],
[en3, direction,
[en3.1, [towards, prep],
[en3.2, [screen, n],
[en3.3, [the, art]]]]])
fr_e([fr1, [tendit, v],
[fr2, dir-obj,
[fr2.1, [bras, n],
[fr2.2, [les, art]]]],
[fr3, direction,
[fr3.1, [vers, prep],
[fr3.2, [ecran, n],
[fr3.3, [l', art]]]]])
clinks ([[en1, fr1], [en2, fr2], [en3, fr3]])

The second source node has two different dependency relations: direct object and possession. The dominant one is a direct object, applied to a commun noun which impose a possession relationship on the Genitive of the personal pronoun. The entire phrase is building in association with the verb, the translatable unit *extended her arms*. The English sub-tree respects the criteria of linear position of wordforms, while the equivalent French sub-tree is more dependent on the criteria for the specific type of direct-object relation.

3. Generalization of the translation examples

3.1. Preliminary

The main problem of the example-based machine translation is the necessity to use a translation example for more than one input situation. The first step is, usually, responsible for the identification of a match between the input lexical string or its sub-

strings, and the translation examples stored in database. Sometimes, this may cause frustration about the translation quality, because the database is not able to manipulate the linguistic flexibility.

One possible solution, for this inconvenience, is given by the combination of semantic-syntactic relations, so that a lexeme from the translation example be able to open different instances for other lexemes, in order to generate the synonymic relation.

The algorithm supposed by this solution has as principal steps to deflect the wordforms from the input sentence, to find all the lexical base forms, to disambiguate them and to establish only one morpho-syntactic value, and-respectively- **to search matches in the example database**. When a verb correspondence is not found, the program searches for terms accepted by the given verb in paradigmatic associations, organized in a semantic relations database. For these situations when there is at least one term, the program checks again for matches and accepts only the trees or sub-trees corresponding to translation input.

Let's take the following syntactic structure extracted from Orwell's corpus, to be translated:

Was preaching freedom of speech.

After the deflection and disambiguation steps, which are not the subject of this paper, the algorithm has to consider the match search. In the example database, the verb to *preach* does not collocate with the noun phrase *freedom of speech*. But this noun phrase has a syntactic dependency relation with another verb, *to advocate*, in the syntactic structure: *was advocating freedom of speech*, with the translation equivalent: *défendait la liberté de parler*. That is why the program is checking now for a semantic relation between the verbs *to preach* and *to advocate*.

In this way, another semantic class 37.1 is found, named- *Verbs of Transfer of a Message*, a class which instantiates for the English language the verb lexemes *preach:2, advocate:2*, and for the French language, the equivalents prôner*:1 and défendre:3*. The search is stopped and the program validates the match between *was preaching freedom of speech* and *was advocating freedom of speech*. The synonymous relations for English are created by means of the lexical semantic ontology WordNet, while for French by using an impressive linguistic resource- TLF, and for Romanian by consulting Luiza and Mircea Seche's dictionary of synonyms.

3.2. The role of the semantic properties of verbs

There is a strong correlation between "the semantic properties of a verb and its syntactic properties, and it seems obvious that speakers can sometimes exploit this pattern to predict form from meaning" (Gropen, J. et al., 1991: 153-195).

This is, in fact, the reason for adding the semantic properties of a verb. A verb that governs a syntactic dependency relation is not isolated in all the verbs worlds, but it is the actant of a synonymy relation with other verbal lexemes. Not all the meanings of a verb participate to create the synset; only those which are grouped around a common meaning.

Indicate the type of the syntactic dependency relation and to create the synsets grouped into verb classes may stimulate the translation quality, because of a better flexibility to the language nature and to its semantic-syntactic representation.

The semantic properties and the syntactic properties are shown in the following translation example:

dădea o muzică stridentă, militărească→ changed over to strident military music→ s'était changée en une stridente musique militaire
ro_e ([ro1, [Verbe de Transformare-> da:9, transmite:13], [da, v],
[ro2, ob-dir,
[ro2.1, [muzică, n],
[ro2.2, [o, art]],
[ro2.3, [stridentă, adj]],
[ro2.4, [militărească, adj]]]]])
en_e ([en1, [had, aux]],
[en2,
[en2.1, [Turn Verbs -> change over:2, convert:2], [change, v],
[en2.2, jonctiv, [over, prep]],
[en3, ob-prep,
[en3.1, [to, prep],
[en3.2, [music, n],
[en3.3, [a, art]],
[en3.4, [strident, adj]],
[en3.5, [military, adj]]]]]])
fr_e([fr1, [etait, aux]],
[fr2,
[fr2.1, [Verbs de Transformer -> changer: 3, transformer: 2], [changer, v],
[fr3, dir-obj, [se, pron]],
[fr4, prep-obj,
[fr4.1, [en, prep],
[fr4.2, [musique, n],
[fr4.3, [une, art]],
[fr4.4, [stridente, adj]],
[fr4.5, [militaire, adj]]]]]])
clinks([[ro1, en2], [ro2, en3]], [[en1, fr1], [en2,fr2], [en3, fr4]])

In conclusion, creating dependency trees means describing for the main verb the semantic class, the associated synset and the types of dependency governed by the verb. In this way, possibilities are generalized to match the input string and the examples is the database, but at the same time a filter is generated from the point of view of syntactic dependency relations. The program has to select only the candidates that governs the same syntactic dependency from the matching candidates set. Together, the semantic-syntactic descriptions have an important role in the translation disambiguation.

4. Conclusions

This paper proposes an approach about the applicability of a translation example database. In order to develop it, I started from the premise of semantic-syntactic relations between two word forms, by following two main ideas: utility and generalization. We consider both the syntactic dependency relations between the verb

and the other parts of speech, and the synonymous relation between a given verb and the lexemes in the same synset, respectively, the same verb class.

At first sight, a translation example database which crammed with this kind of information could be difficult to manipulate, because of the size of search numbers. In order to diminish this inconvenience, I consider three criteria of examples selection before creating a database. First, the very frequent expressions and structures should be considered for both the source and target language. Once this core is realized, the linguist should add the propositional sequences whose meaning is different than the composition of every meaning of its constituents. Finally, the verb phrases from a large corpora, namely "1984", should be added, too.

If the conditions of frequency, of semantic-syntactic-based structures are fulfilled, the search number in database will be diminished. At the same time, there will be a better possibility to identify in the most frequent verb set the ones in a synonymous relation with the input verb form.

The semantic-syntactic disambiguation represents another advantage of designing the example database by the means of the synonymous relation, filtred out by syntactic dependency relations. There are verbs which have more than one meaning, some of them being responsible for the creation of differrent dependency frames. When the synonymy between the meanings of two verbs is evaluated by having the same types of syntactic dependency, the program can identify a certain meaning from the candidates set of the input verb. The operation is also available for the arguments selected by the verb: if the verbs build a synset and one has a syntagmatic line, which is already known, the others imitate its syntactic behaviour.

References

1. Gropen, J., Pinker, S. and R. Goldberg. 1991. "Affectedness and direct object: The role of lexical semantics in the acquisition of verb argument structure", in *Cognition*, 41, 1991, pp. 153-195.
2. Levin, B. 1993. *English Verb Classes and Alternations- A Preliminary Investigation*, Chicago: The University of Chicago Press.
3. Kaji, H., Kida, Y. and Y. Morimoto, Y. 1992. "Learning translation templates from bilingual text", in *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, pp. 672-678.
4. Mel'cuk, I. 2003. "Levels of Dependency in Linguistic Description: Concepts and Problems", in *Dependency and Valency. An International Handbook of Contemporary Research*, in (V. Agel, L. Eichinnger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin) (eds.), vol. 1, Berlin - New York, W. de Gruyter, pp. 188-229.
5. Nagao, M. 1984. "A framework of a mechanical translation between Japanese and English by analogy principle", in *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, Lyon, France, pp. 173 – 180.
6. S. Sato And M. Nagao. 1990. "Toward memory-based translation", in *Proceedings of the 13th conference on Computational Linguistics, v*ol. 3, Helsinki, Finland, pp. 247 – 252.