Cross-Language Transfer Of Syntactic Relations Using Parallel Corpora

Verginica Barbu Mititelu
Romanian Academy Research Institute for Artificial Intelligence,
13 Septembrie, 13, 050711, Bucharest
vergi@racai.ro

Radu Ion
Romanian Academy Research Institute for Artificial Intelligence,
13 Septembrie, 13, 050711, Bucharest
radu@racai.ro

Abstract

The data that we present in this paper are only some preliminary results of an experiment that we are carrying out in order to test the feasibility of the automatic cross-lingual transfer of syntactic relations using parallel corpora annotated at the morphological and syntactic level. The aim of our experiment is to test whether it is possible (and also to what degree) to automatically transfer syntactic relations (as they are lexicalized in a corpus) from a resource-rich language into another language with fewer resources, using parallel corpora.

1 Introduction

Natural Language Processing (NLP) applications make great use of linguistic resources. However, the development of such resources is time- and money-consuming.

Lately, the NLP community has started using alternative strategies for getting the necessary resources. One such strategy is the use of knowledge in one language to help solving tasks in another language. One example of knowledge transfer is to take advantage of the resources built for one language to induce knowledge in a resource-poor language. This is made possible by the existence of aligned parallel corpora.

What we present below are some preliminary results of an experiment in which we test the possibility of automatic transfer of syntactic relations from a resource-rich language (English) into a resource-poor one (Romanian).

2 Assumption

We started from the Direct Correspondence Assumption (DCA) as it is formulated in (Hwa et al. 2002b):

> Given a pair of sentences E and F, that are (literal) translations of each other with syntactic structures Tree$_E$ and Tree$_F$, if nodes x_E and y_E of Tree$_E$ are aligned with nodes x_F and y_F of Tree$_F$, respectively, and if syntactic relationship $R_{(xE \text{ and } yE)}$ holds in Tree$_E$, then $R_{(xF \text{ and } yF)}$ holds in Tree$_F$.

As formulated, DCA is applicable to parallel treebanks. However, we have modified it as follows, so that it serves our purpose:

> Given a pair of sentences E and F, that are (literal) translations of each other, if words x_E and y_E of E are aligned with nodes x_F and y_F of F, respectively, and if syntactic relationship $R_{(xE \text{ and } yE)}$ holds in E, then $R_{(xF \text{ and } yF)}$ holds in F.

The reformulated DCA ensures the cross-lingual transfer of syntactic relations existent between two lexical items into the same syntactic relations between the translation equivalents of those lexical items in a parallel corpus.

3 Resources and Tools

The languages in focus here are English and Romanian. The parallel corpus that we use is George Orwell’s *1984*, which was developed during the MULTEXT-EAST project (Dimitrova et al. 1998). This corpus is rather small, as one can see in Table 1:

	English	Romanian
Translation units	6411	
Unique lemmas	7359	7248
Unique word forms	10152	15112

Table 1. Quantitative data about the corpus.

1984 is XML encoded obeying a simplified form of the XCES standard (Ide et al. 2000) and is sentence aligned, tokenized and morpho-syntactically annotated with a tagset that is specially designed so as to encode the morpho-syntactic values in a language independent way. Besides the above-mentioned annotation, we have also used a simple regular expression chunker to mark the constituents of a given sentence: noun phrases, prepositional phrases, adjectival and adverbial groups and verbal groups. Two separate grammars have been written (one for English and the other for Romanian) that generate PERL regular expressions over sequences of POS tags of English and Romanian for each type of phrase. For instance, “the big lazy dog” and its corresponding Romanian translation “câinele cel mare şi leneş” are tagged with the noun phrase (NP) chunk label. Every phrase has its own ordering label in the sentence so that in “The cat with the furry tail took the fish from the bowl.”, “the cat” receives the NP1 chunk label, “the furry tail” NP2 and so on. For this particular sentence, our chunker offers the following annotation:

(NP1 The cat)NP1
(PP1 with
(NP2 the (AP1 furry)AP1 tail)NP2
)PP1
(VP1 took)VP1 (NP3 the fish)NP3
(PP2 from (NP4 the bowl)NP4)PP2.

1984 has also been word aligned using a combined word aligner (COWAL) (a program that for every index of a word in a source language sentence gives the index of a word in the target language sentence to which the source word aligns) described in (Tufiş et al. 2005). The above-mentioned chunks were successfully used in reducing the ambiguities that a word aligner has to face assuming that in most cases, a chunk in Romanian aligns with other chunk(s) in English.

For the transfer of the syntactic relations between English and Romanian, another annotation of the English part of *1984* was available: the syntactic analysis using a functional dependency grammar (FDG). In particular, we had at our disposal the output of the FDG parser described in (Tapanainen and Järvinen, 1997a) and (Tapanainen and Järvinen, 1997b) on the English version of *1984*. The dependency between two words is marked by specifying the index of the governor and the function name at the dependant position as in Figure 1 (where 0 represents the root of the syntactic tree):

IDX	WORDFORM	FDG ANNOTATION
0		
1	It	subj:2
2	was	main:0
3	a	det:6
4	bright	attr:5
5	cold	attr:6
6	day	tmp:2
7	in	tmp:6
8	April	pcomp:7
. . .		

Figure 1: Representation of dependencies.

The FDG parser was applied on an older version of the annotated English *1984* and due to the fact that the tokenization of this version and the tokenization of our current version of *1984* are different, we were forced to use a subset of translation units from our corpus so that the following conditions held:

1. The selected translation unit contains sentences that are in a 1:1 correspondence, meaning that the English sentence is translated by a single Romanian sentence. This restriction is imposed because of the different outputs of the COWAL word aligner and the FDG parser with respect to indexing. While the FDG parser resets the numbering of words with the beginning of each sentence, COWAL considers two or more sentences as one and as such, the indexes are consecutive;
2. The tokenization of the English sentence from our corpus is identical with the tokenization of the same sentence in the FDG annotated version.

After making this selection, there were 1537 translation units left in which only 1:1 sentence alignments exist. This selection also favors COWAL because the shorter the sentences, the better the word alignment accuracy.

4 Problems for DCA

The parallel corpus that we have made use of raises some problems due to the strategy adopted by the translator: his/her aim was to give a literary translation of Orwell’s novel, not a literal one which keeps as close as possible to the original version. One such problem is posed by the fact that the translator chose to introduce in the Romanian text some verbs that were not present in the English version, in order to be able to render the exact meaning from the original. Consider the following equivalent sentences as an example:

(1) En: He crossed the room into the tiny kitchen.
Ro: Traversă camera şi merse în bucătărie.

The translation equivalence pairs extracted by the COWAL word aligner are the following: *crossed*-- *traversă*, *the room*--*camera*, *into*--*în*, *the kitchen*--*bucătărie*, .--.. As it can be seen, the Romanian *merse* lacks a lexicalized, although semantically implicit, English equivalent in the respective sentence[1].

A more problematic case is the following:

(2) En: He moved over to the window: a smallish, frail figure, the meagerness of his body merely emphasized by the blue overalls which were the uniform of the Party.
Ro: Winston se duse către fereastră: avea o figură fragilă, mai degrabă mică, iar salopeta albastră, care era uniforma Partidului, scotea în evidență cât era de slab.

The equivalent pairs are: *moved*--*se duse, to*--*către*, *the window*--*fereastră*, *:*--*:*, *a*--*o*, *smallish*--*mică*, *,*--*,*, *frail*--*fragilă*, *figure*--*figură*, *,*--*,*, *the overalls*--*salopeta*, *blue*--*albastră*, *which*--*care*, *were*--*era*, *the uniform*--*uniforma*, *the Party*--*Partidului*, .--..

The problems here are due to the introduction of the verb *avea* in Romanian and to the fact that a passive participial construction is translated into one with the verb in the active voice; the translator avoided using the Romanian noun *slăbiciune* corresponding to *meagerness*, as the former has a wide-spread abstract meaning, instead of sending the reader's mind to the fact that the body is thin. The use of the Romanian adjective *slab* makes the translation of *body* useless, as *slab* is used, with its first meaning, to refer to persons' bodies.

We decided not to take into consideration sentences as (1) to (2) above, at least in the first part of our research, as they do not seem relevant for the task of transferring syntactic functions from one language into another.

The sentences we focused on for the transfer are those where the translator kept close to the original both semantically and syntactically, trying to use the most appropriate Romanian equivalents of the English words, and also in similar syntactic structures.

While making the selection of the units that are worth taking into consideration for our task, we manually corrected the alignments that were wrongly identified by the COWAL aligner. As previous similar experiments proved (Hwa et al. 2002a), the quality of the alignment results influences the quality of the syntactic transfer.

5 The transfer procedure

Having ensured that the English sentence is translated as closed as possible into Romanian with respect to the syntactic realization of its content, we pursued the following steps for every syntactic dependency relation ($\mathtt{srel}$) in the English sentence:

1. extract the alignment indexes of the governor and dependent of the English $\mathtt{srel}$ in Romanian. We thus obtain two indexes sets: $\mathtt{G(ro)}$ and $\mathtt{D(ro)}$;
2. if $\mathtt{|D(ro)|}$ is equal to $\mathtt{|G(ro)|}$ and equal to 1 ($|\cdot|$ being the set cardinality function) and if $\mathtt{d(ro)} \in \mathtt{D(ro)}$ and $\mathtt{g(ro)} \in \mathtt{G(ro)}$, and $\mathtt{d(ro)} \neq \mathtt{g(ro)}$, then transfer the relation $\mathtt{g(ro)\ srel\ d(ro)}$ (we simply transfer the relation from English to Romanian provided that "both ends of the (relation) arrow" point to single (different) indexes in Romanian);
3. if either $\mathtt{|D(ro)|>1}$ or $\mathtt{|G(ro)|>1}$, we employ a rule-based algorithm for the extraction of the group head from the Romanian alignment indexes set that has more that one index in it. For instance, if the alignment *went*--*se duse* is encountered, one such rule extracts the Romanian verb *duse* as the head of the construction (the index of which comprises the new, reduced set, $\mathtt{D(ro)}$ or $\mathtt{G(ro)}$) and the transfer algorithm continues with step 2.

[1] For the inexistence of an equivalent for the English *he*, see below the discussion about the pro-drop phenomenon.

Table 2 gives the percentages of the relations[2] transferred in Romanian from the total number of relations present in the English part of the bitext. We assume, in concordance with the DCA, that the higher the transfer percentage, the more chances there are for the relation to hold in Romanian, as well.

In addition to these relations, we discovered some relations (see the LOST column) that are, in some cases, English syntax tailored. That is, at step 2 in the transfer algorithm, if d(ro)=g(ro), the relation is lost (because the "relation arrow" will start and point to the same index in Romanian). For instance, we see that the relation det was lost 173 times. That is because in Romanian the definite article is placed as a suffix on the determined word, while in English, it is placed before the determined word being itself a lexeme. So, in *the car*--*maşina* the relation det between *car* and *the* is lost in Romanian. Obviously, these relations were not transferred.

Rel	RO	Lost	EN	%
pth	1	0	1	100%
pccomp	1	0	1	100%
qn	10	0	12	83.33%
agt	4	0	5	80%
neg	10	0	13	76.92%
oc	3	0	4	75%
dat	3	0	4	75%
cnt	8	0	11	72.72%
ad	25	0	35	71.42%
pcomp	218	9	316	68.98%
sou	6	0	9	66.66%
loc	26	0	39	66.66%
meta	40	0	63	63.49%
comp	70	1	112	62.5%
attr	151	4	245	61.63%
cc	94	2	155	60.64%
pm	44	1	75	58.66%
obj	79	2	137	57.66%
mod	114	1	201	56.71%
ha	41	0	74	55.4%
cla	8	0	15	53.33%
tmp	23	0	46	50%
man	16	0	32	50%
goa	7	0	14	50%
subj	121	2	319	37.93%
frq	8	0	22	36.36%
det	126	173	355	35.49%
dur	1	0	3	33.33%
copred	1	1	3	33.33%
cnd	1	0	4	25%
v-ch	35	48	143	24.47%
phr	3	0	15	20%
ins	0	0	1	0%

Table 2: Percent of transferred relations.

6 Comments on the transfer possibilities

6.1 Perfect transfer

Some preliminary results showed us that the cross-lingual transfer of syntactic relations is possible most of the times, thus confirming the DCA:

(3) En: The hallway smelt of boiled cabbage and old rag mats.
Ro: Holul blocului mirosea a varză călită şi a preşuri vechi.

The equivalents are: *the hallway*--*holul*, *smelt*--*mirosea*, *of*--*a*, *cabbage*--*varză*, *boiled*--*călită*, *and*--*şi*, *mats*--*preşuri*, *old*--*vechi*, .--.. Although the Romanian sentence has an extra word, *blocului* "of the block", and the English one, in its turn, has one extra word, *rag*, these do not affect the verbal complementation, which is the same in both sentences: subj (between *hallway* and *smelt*, respectively between *holul* and *mirosea*), phr[3] (between *of* and *smelt*, respectively between *a* and *mirosea*).

6.2 Transfer with some amendments

Sometimes, although the syntactic structures in the two languages are not similar, some relations can be transfered. It is the case of the "active" constructions in English which are translated into Romanian with their "passive" counterparts:

(4) En: It was partly the unusual geography of the room that had suggested to him the thing that he was now about to do.
Ro: Lucrul pe care avea de gând să-l facă îi fusese sugerat, în parte, de această geografie neobişnuită a camerei.

[2] For the description of these relations, as well as for examples, one can see (Tapanainen and Järvinen 1997b).

[3] The verb-particle relation (for short phr) is the relation holding between a verb and its particle. (Tapanainen and Järvinen 1997b)

Had suggested is in the active voice and contracts the following relations: subj (with *that*), dat[4] (with *to him*), and obj[5] (with *the thing*). Its Romanian counterpart, *fusese sugerat*, establishes the dat relation with *îi* (the equivalent of *him*), the subj relation with *lucrul* (the equivalent of *the thing*) and phr relation with the group headed by the preposition *de*. From the morpho-syntactic annotation we can get the information that the Romanian sentence is in the passive voice, so we can create a rule for the transfer of syntactic functions, a rule which may help the conditioned "inverse" transfer of some functions: the subj is transfered as an obj, and the obj as a subj.

Another example of transfer with amendments is illustrated by the following example:

(5) En: It was a peculiarly interesting book.
Ro: Era o carte deosebit de frumoasă.

The equivalence pairs are: *was*--*era*, *a*--*o*, *peculiarly*--*deosebit*, *interesting*--*frumoasă*, *book*--*carte*, . --..

Such cases are rather frequent: Romanian lacks an equivalent for the English dummy (anticipatory) *it*, so the subject relations existing between *it* and *was* has no Romanian counterpart. However, the comp relation[6] existing between *book* and *was* cannot be transfered as such in Romanian: the relation holding between *carte* and *era* is subj. A further step in our attempt to automatically transfer syntactic functions would be the appropriate transfer of such functions via their transformation into the correct ones in the target language (this involves the transfer of the comp relation as subj in the target language).

6.3 Language specific phenomena

The typological differences between the two languages considered make idiosyncrasies unavoidable.

Unlike English, Romanian is a *pro-drop* language[7], thus many subj relations from the source language remain without an equivalent in the target one. Consider the following example, where the Romanian sentence lacks a lexicalized subject for the verb *erai*, the equivalent of *had*:

(6) En: You had to live.
Ro: Erai obligat să trăieşti.

Another characteristic of Romanian is the doubling phenomenon: a direct or indirect object lexicalized as an NP with some semantic and/or syntactic characteristics (Guțu Romalo 1973) is obligatorily doubled by a pronominal clitic with which it shares the grammatical information of case, gender, person, and number. In the parallel corpus one can find an English equivalent only for the NP in such cases and the relation can be safely cross-lingually transferred.

(7) En: He had set his features into the expression of quiet optimism which it was advisable to wear when facing the telescreen.
Ro: Îşi compusese pe față acea expresie de optimism liniştit pe care era indicat să o abordezi când stăteai cu fața la tele-ecran.

Which is aligned with *care*, this one being doubled by *o*, which lacks an English counterpart. The obj relation between *which* and *wear* can be transfered between *care* and *abordezi* (the equivalent of *wear*), while between *o* and *care* an anaph relation can be established, after resolution.

A further step (at the target language level only) would be taking a decision concerning the treatment of the clitics in such situations. The possibilities would be either to treat them at the morphological level, so part of the verbal morphology, or to treat them at the syntactic level and postulate a language-specific relation (which we may call anaph) holding between the clitic and its co-referent NP. The grammatical information shared by the two would ease the resolution.

6.4 Impossibility of transfer

Besides such idiosyncrasies due to the typological differences between the chosen

[4] This is the relation established between the indirect object (in Dative) and the verb whose argument it is.

[5] This relation is established between the verb and its object. According to Tapanainen and Järvinen (1997b) the notion of object comprises essentially all types of second arguments, except for subject complements.

[6] Tapanainen and Järvinen (1997b) establish this relation between the copular verb and the subject complement, namely the second argument of a copular verb.

[7] Pro(noun)-drop(ping) languages are languages where pronouns can be deleted when (grammatically or pragmatically) inferable. Romanian allows deletion of subject pronouns.

languages there are also cases when the equivalent verbs display a different syntactic behavior.

(8) En: I like to see them kicking.
Ro: Îmi place să-i văd dând din picioare.

Like takes a `subj` (*I*) and an `obj` (*see*), while *place* is involved in a `dat` (*îmi*) and an `obj` (*văd*) relations.

For some English adverbs and adjectives Romanian has a prepositional construction as equivalent, as below:

(9) En: The girl with dark hair was sitting immediately behind.
Ro: Fata cu părul negru stătea exact în spate.

The adverb *behind*, being in `ha`[8] relation with the verb *sitting*, is the equivalent of the Romanian prepositional phrase *în spate*. If we choose to link these two words (*în_spate*), then the problem disappears: there remains a 1:1 equivalence between the two and the relation can be safely transfered.

7 Conclusions and further work

The preliminary results of the syntactic annotation transfer justifies our belief that an automatic procedure of transferring syntactic relations in Romanian is reliable provided that all resources are present with the required level of annotation. However, language specific structures and grammatical phenomena require the pre- and post-processing of the data. That is why, our very next step is the implementation of linguistic rules for eliminating the noise obtained after the transfer.

We are perfectly aware that our corpus is too small[9]. As we needed a corpus as well as possibly aligned at the word level, we restricted our analysis to a limited number of sentences for which we could manually check the results of the COWAL aligner. For the future, as we will have a better version of the COWAL, we will be able to extend the corpus. Through this study we aim at enriching the Romanian WordNet (Tufiş et al. 2004) developed during the BalkaNet project with the verb frames extracted by word alignment and syntactic relation transfer. This could eventually enable the development of a Romanian parser, which could, in turn, enable other important NLP applications.

[8] This is a default category to attach post-verbal adverbs and prepositional clauses to the verbal element (Tapanainen and Järvinen 1997b).

[9] That is too small to be sure that all the syntactic relations from English had a fair chance of being tranferred. We are also aware that the resulting (Romanian) corpus is not (yet) to be used as training data.

References:

L. Dimitrova, T. Erjavec, N. Ide, H. Kaalep, V. Petkevic, D. Tufiş. 1998. *Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages*, COLING, Montreal.

Valeria Guţu Romalo. 1973. *Sintaxa limbii române. Probleme şi interpretări*, Bucharest, Editura Didactică şi Pedagogică.

Rebecca Hwa, Philip Resnik, Amy Weinberg. 2002a. Breaking the Bottleneck for Multilingual Parsing. In *Workshop on "Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data"*, Third International Conference on Language resources and Evaluation (LREC-2002), Las Palmas, Canary Islands, Spain.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002b. Evaluating Translational Correspondence using Annotation Projection. In *Proceedings of the 40th Annual Meeting of the ACL*. Philadelphia, PA.

N. Ide, P. Bonhomme, L. Romary. 2000. XCES: An XML-based Standard for Linguistic Corpora. In *Proceedings of LREC2000, Athens, Greece*, 825-30.

Pasi Tapanainen and Timo Järvinen. 1997a. *A non-projective dependency parser*. In the *Proceedings of the 5th Conference on Applied Natural Language Processing* (ANLP'97), ACL, Washington, D.C., 64-71.

Pasi Tapanainen and Timo Järvinen. 1997b. *A dependency parser for English*. Technical Report no. TR-1, Department of General Linguistics, University of Helsinki, Finland.

Dan Tufiş, Eduard Barbu, Verginica Barbu Mititelu, Radu Ion, Luigi Bozianu. The Romanian Wordnet. In *Romanian Journal on Information Science and Technology*, Dan Tufiş (Ed.) Special Issued on BalkaNet, Romanian Academy, vol7, no. 2-3, 2004, 105-122.

Dan Tufiş, Radu Ion, Alexandru Ceauşu, Dan Ştefănescu. 2005. Combined word alignments. In *Proceedings of the ACL 2005 Workshop on Parallel Text*, Ann Arbor, Michigan, USA.