PAPER No. 826

Tagging of very large corpora: Topic-Focus Articulation

Abstract

After a brief characterization of the theory of the topic-focus articulation of the sentence, rules are formulated that determine the assignment of appropriate values of the TFA attribute in the process of semantico-syntactic tagging of a very large corpus of Czech.

1 Introduction: The Prague Dependency Treebank (PDT)

PDT is a corpus (a part from the Czech National Corpus), tagged on the following levels:

(i) morphemic (POS and annotations using a very large number of tags, as required by the language with rich inflection; cf. [1]);

(ii) 'analytic' (dependency syntax, with nodes for all word occurrences, also for punctuation marks etc., and with the tags for morphemic units and for basic kinds of surface syntactic relations (Subject, Object, Adverbial, Adjunct), cf. [2]

(iii) tectogrammatical (underlying) syntax, with a much more detailed classification of syntactic relations and with nodes for autosemantic lexical occurrences only (with indices corresponding to the relations and to morphological values such as Preterite, Conditional, and also as the prototypical values of 'in', 'into', 'on', 'from', etc.).

2 Representing Topic-Focus Articulation (TFA) in TGTSs

2.1 A brief characterization of TFA

The tectogrammatical tree structures (TGTSs) should capture not only the syntactic (dependency) relations, but also the TFA of the utterances in the corpus, since TFA is expressed by grammatical means and is relevant for the meaning of the sentence (even for its truth conditions), i.e. it constitutes one of the basic aspects of underlying structures. The semantic relevance of TFA can be illustrated by examples such as (1), which is a translation of the Czech ex. (1') (the capitals denote the placement of the intonation centre, i.e. the focus proper):

(1)(a) English is spoken in the SHETLANDS.

(b) In the Shetlands, ENGLISH is spoken.

(1') (a) Anglicky se mluví na Shetlandských OSTROVECH.

(b) Na Shetlandských ostrovech se mluví ANGLICKY.

The communicative function of the sentence can basically be rendered by understanding its topic (T) as "what is the sentence about", and its focus (F) as the information that is asserted about the topic, i.e., schematically, the interpretation of the sentence S can be understood as

S = F(T)

Thus, (1)(a) asserts, on its preferred reading (with just the locative modification constituting its focus) about where English is spoken that it is in the Shetlands, which hardly can be accepted as true w.r.t. what we know of the actual world, if no specific context is present. (1)(b) is understood as true, stating about E. that it is spoken in the S.

In the TGTSs the order of nodes is such that all parts of T precede all parts of F.

Moreover, the order of nodes corresponds to the scale of communicative dynamism (CD, see Section 3 below); a less dynamic node prototypically has the broader scope than a more dynamic one (if the nodes correspond to operators). F proper is then the most dynamic (the rightmost) node.

TFA is relevant also for the semantics of negation:

(2) John didn't come because he was ILL.

(a) The reason for John's not-coming was his illness.

(b) The reason for John's coming (e.g. to the doctor) was not his illness but something else (e.g. he wanted to invite the doctor for a party).

With the paraphrase (a), the negated verb 'come' is included in T, i.e. the fact that John's being ill is the cause of an event is asserted about the event that he did not come. With (b), the main verb 'come' also belongs to T, but what is negated, is the relation between T and F: John came, but what is asserted about his coming is that the cause of this event was not his illness (he might have been ill, though).

Every node in a TGTS is either contextually bound (CB) or non-bound (NB); this opposition is a linguistic counterpart of the cognitive dichotomy of 'given' vs. 'new', where also an item presented as occupying a newly characterized specific position (often in relation to one or more 'given' items) has the feature 'new', cf.:

(3) Give this to my MOTHER.

(4) (Mary knows both Peter and Jane.) However, this time she only invited HER.

As shown by the presence of the indexical pronoun in (3) and of a (although stressed) anaphoric pronoun in (4), neither of the referents can be interpreted as not 'given', or at least 'known' in some sense. Even so, in these examples, both 'mother' and 'her' occur as NB; their stress indicates their function as F proper of the respective sentence.

Prototypically, an NB node belongs to F and a CB node is in T; however, a node not dependent immediately on a finite verb (esp. an adjunct) need not meet this condition. Thus, in (5), 'my' as a shifter, directly determined by the conditions of the discourse, is CB, although belonging to F, since it depends on a part of F (see [3] for a definition of T and F on the basis of contextual boundness and of syntactic dependency, as well as for other details of the given descriptive framework).

2.2 The attribute TFA in PDT:

Three values of the attribute TFA are distinguished with every node in a TGTS:

(i) T – a non-contrastive CB node, which always has a lower degree of CD than its governor, if any;

(ii) F – an NB node (if different from the main verb, then following after its head word in the TGTS)

(iii) C – a contrastive CB node

Examples:

(5) (Volby v Izraeli.)
Po volbách(T) si Izraelci(T) zvykají(F) na nového(F) premiéra(F).

(Headline in the newspapers: Elections in Israel.)
After the elections(T), the Israelis(T) get used(F) to a new(F) Prime Minister(F).

(6) Sportovec(C) on(T) je(F) dobrý(F), ale jako politik(C) nevyniká(F).

Sportsman(C) he(T) is(F) good(F), but as a politician(C) he does not excel(F).

The instructions for the assignment of the values of TFA can be briefly specified as follows, if the surface word order and the position of the (typically falling or rising-falling) sentence stress (intonation center, IC) is taken into account, as well as the 'systemic (canonical) ordering of the kinds of dependents (which, in fact, can differ with different head words; it is specified either in the valency frames in the individual

lexical entries, or, if possible, by means of indices concerning lexical classes and subclasses):

(i) the bearer of IC => F (typically = the rightmost dependent of the verb)

(ii) if the IC is placed on a node other than the rightmost one, the complementations placed after IC => T

(iii) a left side dependent of the verb => T or C (except for cases in which it clearly carries IC)

(iv) the verb and those of its dependents that stand between the verb and the F-node (see (i)) and that are ordered (without an intervening sister node) according to the systemic ordering (SO) => F

Note: For Czech, the SO of the main types of dependency has been found (on the basis of empirical analysis of texts and of experiments with groups of speakers) to have (with most verbs and other heads) the following form, as for the main kinds of dependents:

Actor < Temporal < Location < Instrument < Addressee < Patient < Effect

(v) embedded attributes => F (unless they are only repeated or restored)

(vi) indexical expressions ('já' [I], 'ty' [you], 'teď' [now], 'tady' [here], weak forms of pronouns, pronominal expressions with a general meaning ('někdo' [somebody], 'jednou' [once upon a time]...) => T (except in cases of contrast or as bearers of IC)

(vii) strong forms of pronouns => F (after prepositions, the assignment of T or F in Czech is guided by the general rules (i) through (iii))

(viii) restored nodes, deleted in the surface forms of sentences => T

Note: There are special cases of coordination, both in Czech and in English, which do not meet this condition: e.g. in 'They drank white and red wine' the first occurrence of 'wine', which may be NB, is deleted.

(ix) a node N dependent to the left in a way not meeting the condition of projectivity: => C (this node is then placed more to the right, to meet that condition)

(x) the nodes depending on N (directly or indirectly) will move together with N: => T or F (according to the rules above)

Note: The resulting TGTSs are projective, i.e. for every pair of nodes x, y in a TGTS it holds that if x depends on y and x follows (precedes) y, then every node z following (preceding) y and preceding (following) x is subordinate to y, where 'subordinate to' is a transitive closure of 'dependent of'. Thus, 'not to meet the condition of projectivity' concerns the 'analytic' trees; this means, in other words, that this condition would not be met if the positions of x and y in the left-to-right order of the nodes in the TGTS (in the 'underlying word order') corresponded to their positions in the surface (morphemic and 'analytic') word order.

Example (with a very simplified linearized notation of the TGTS, in which every dependent is closed in its pair of parentheses):

(7) K jásotu(C) není(F) nejmenší(F) důvod(F).
For triumphing is-not the-least reason

(7') (neg.F) být.F (důvod.F (jásot.C) (nejmenší.F))
(neg.F) be.F (reason.F (triumphing.C) (least.F))

A sentence with a non-prototypical placement of the IC:

(9) (Většina ministrů Stěpašinovy nové vlády patří k věrným druhům nejznámějšího ruského intrikána Berezovského.)
(The majority of the ministers of Stěpašinov's new government belongs to

faithful friends of the best known Russian intriguer Berezovskij.)

I(F) AKSJONĚNKO(F) udržuje(T) s
Even(F) AKSJONENKO(F) keeps(T) with
Berezovským(T) blízké(F) styky(T).
Berezovskij(T) close(F) contacts(T).

3 The special case of focus sensitive particles

Since the focus sensitive particles are identified (by the word-class value RHEM for 'rhematizer' or 'focalizer'), it is possible to use PDT also for a specification of their occurrences (a) in different positions in sentence structure and (b) in TFA. The starting hypotheses, which might be checked on the basis of PDT, are as follows:

3.1 Focus sensitive particles in prototypical positions

The prototypical syntactic position of a focalizer is that of a dependent of a verb node; thus, in examples like (10) or (11), it is possible to specify the scope of the focalizer as the whole subtree subordinated to the verb (where 'subordinated' is understood as the transitive closure of 'dependent', in the reflexive sense, so that the verb itself is included); the scope is divided into background and focus of the focalizer (ff), as will be specified in 3.2. Thus, in the interpretation of (10) on the reading represented (with many simplifications) by (10') it is included that (according to what P. knows) among those whom T. saw there was no one else than M (i.e. while 'T. saw' constitutes the background of 'only', its ff is 'Mary'). Similarly, if in (11) the negation (although expressed by a prefix in Czech) is handled as a dependent of the verb, its background is the subject and ff includes both the verb and the object.

(10) Pavel ví, že Tomáš viděl jen MARII.
'Paul knows that Thomas saw only MARY.'

(10') (Paul) knows ((Thomas) saw (only) (Mary))

(11) Martin nečte NOVINY.
'Martin not-reads NEWSPAPERS.'

In (12) only the adjective constitutes the ff of 'only', its background consisting of 'car' (among all cars, P. only wants a blue one); thus, the focalizer can best be described here as depending on 'car'.

(12) Petr chce jen MODRÉ auto.
'Petr wants only (a) BLUE car.'

3.2 Focus sensitive particles in the hierarchy of communicative dynamism

The primary position of a focalizer in a TR is at the boundary between the topic and the focus of the verb clause and the focus of the clause is then identical to the focus of the focalizer. If a focalizer occupies another position, then its focus contains those items which in the TR are placed between this focalizer and the next superscript c to the right (and thus are more dynamic than the focalizer).

It should be noted that CD is understood here as a partial ordering defined so that:

(i) in every set of a head and its daughter nodes, every daughter node placed to the right of its head is more dynamic than every daughter node placed to the left of its head;

(ii) the relation 'more dynamic' is determined by the irreflexive transitive closure of (i).

Thus, e.g. in the TR (10'), 'knows' is more dynamic than 'Paul' and less dynamic than 'saw' according to (i), and both 'only' and 'Mary', being more dynamic than 'saw', are more dynamic than 'knows' according to (ii); however, 'Thomas' is neither more nor less dynamic than 'knows'.

The underlying word order W (a linear ordering) is then defined on the basis of CD, with (iii) and (iv) holding for every two nodes *x* and *y* in a tree:

(iii) if node *x* is more dynamic then node *y*, then *x* follows *y* under W;

(iv) if node *x* follows node *y* under W, node *u* is subordinated to *x* and node *z* is subordinated to *y* (where 'subordinate to' is the transitive closure of 'dependent on'), then *u* follows both *y* and *z*, and *x* follows *z* under W.

Examples (13) - (16) (most of which we owe to B. H. Partee) point to the possibility to derive recursivity from CD if each degree of dynamism is understood as a subordinated step in the overall topic-focus structure of the sentence and if the syntactic relations are handled as corresponding to covert operators in the TSs. In (13) the focalizer occupies the primary position, in (14) it occurs within the topic, and thus corresponds to the Op in the complex R, in (15) two overt focalizers are present in these two positions; example (16), introduced by J. Jacobs, is reanalyzed here as having the intonation center on the subject and a phrasal stress, denoted here by inverted commas, on the object noun.

(13) John's brother also drank WINE.

(13') ((John's) brother).T (also) drank.F (wine).F

(13") Op also, R x = (Op APPURT, R John's NS brother), NS (Op ASSERT, R x, NS (Op OBJ, R drank, NS wine)

Note: To indicate the syntactic relations in the TSs, we use covert operators corresponding to the tectogrammatical relations and modalities: ASSERT stands for the positive declarative modality of the predicative relation, APPURT for Appurtenance (broader than possession), OBJ for Objective, etc. The symbols T, F and C denote the relevant items as included in the topic, focus and contrastive topic, respectively.

(14) (Has even JERRY noticed anything?) (Well,) even Jerry has noticed Mike's ANNOYMENT.

(14') (even) (Jerry).C has-noticed.T ((Mike's) annoyment).F

(14") Op ASSERT, R (Op even, R (Op OBJ, R x, NS has-noticed), NS Jerry), N x = (Op APPURT, R Mike, NS annoyment)

(15) (What did even Paul realize?) Even Paul realized that Jim only admired MARY.

(15') (Jim).T admired.T (only) (Mary).F

(15") Op ASSERT, R (Op even, R (Op OBJ, R x, NS realized), NS Paul), NS x = (Op ASSERT, R Jim, NS (Op OBJ, R admired, NS y), NS y = Mary)

(16) Sogar PETER kennt nur einen *Roman* von Goethe.
[Even PETER knows only a *novel* by Goethe.]

(16') (nur) ((von-Goethe) einen-Roman.C) kennt.T (sogar) (Peter).F

(16") Op sogar, R (Op ASSERT, R x, NS (Op OBJ, R kennt, NS (Op APPURT, R y, NS Goethe), NS y = Roman), NS x = Peter

4. Summary

After a brief characterization of the Prague Dependency Treebank and of the Praguian theory of Topic-Focus Articulation we have presented a proposal how to integrate into the tagging system capturing the underlying structure of sentences the main aspects of the information structure of the sentence. The last section exemplifies how the proposed approach makes it possible to analyze structures with the so-called focus sensitive operators.

References

[1] Hajič J. and Hladká B. (1997) *Probabilistic and rule-based tagger of an inflective language - a comparison*. In "Proceedings of the Fifth Conference on Applied Natural Language Processing", Washington, D.C., pp. 111-118.

[2] Hajič J. (1998) *Building a syntactically annotated corpus: The Prague Dependency Treebank*. In: "Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová", E. Hajičová, ed., Karolinum, Prague, pp. 106-132.

[3] Hajičová E., Partee B. and Sgall P. (1998) *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer, Dordrecht.