Identifying Topic and Focus by an Automatic Procedure

Eva Hajičová & Petr Sgall
Institute of Formal and Applied Linguistics
Charles University
Malostranské nám. 25, 118 00 Praha 1
Czech Republic
(hajicova@cspguk11.bitnet, sgall@cspguk11.bitnet)

Hana Skoumalová
Institute of Theoretical and Computational Linguistics
Charles University
Celetná 13, 110 00 Praha 1
Czech Republic
(skoumal@praha1.ff.cuni.cs)

Abstract

An algorithm for automatic identification of topic and focus of the sentence is presented, based on dependency syntax and using written input, which is much more ambiguous than spoken utterance.

1. The dichotomy of topic and focus, based, in the Praguean Functional Generative Description, on the scale of communicative dynamism (underlying word order), is relevant not only for a possible placement of the sentence in a context, but also for its semantic interpretation.

The **underlying word order** differs from the surface one especially in that the verb stands more to the right than all its complementations belonging to the topic of the sentence (or to the local topic of the clause headed by the verb), and more to the left than those belonging to the focus. Using a **dependency** grammar (or, more or less equivalently, a flat structure in a constituency based grammar), we can illustrate this by the following example, where (1') is a simplified underlying representation of (1) on a reading answering e.g. the question *Where has Charles found my pen?*:

(1) Charles has found your pen in a box lying on the table.

(1') $(\text{Charles})_{\text{Act}}$ $((\text{you})_{\text{Appurt}}$ $\text{pen})_{\text{Obj}}$ find.Perf $(\text{box.Indef}$ $((\text{Rel})_{\text{Act}}$ lie $(\text{table})_{\text{Loc.on}})_{\text{Gener}})_{\text{Loc.in}}$

In (1') every pair of parentheses encompasses a dependent item (i.e. corresponds to an edge of the linearized dependency tree), the indices of parentheses denote kinds of dependency (valency slots, or theta roles and adjuncts): Act stands for Actor (underlying Subject), Appurt for Appurtenance (Possessivity in a broader sense), Obj for Objective (underlying Object), Loc for Locative, Gener for the General Relationship (of an adjunct to its head); the other indices denote values of morphological categories (Perfect, Indefiniteness) and of adverbial prepositions (*in*, *on*), Rel denotes a relative pronoun (here

deleted on the surface). For more details of the descriptive framework used, see Sgall et al. (1986, Chapters 2 and 3).

An automatic identification of topic and focus may use the input information on surface word order, on the dependency relations between autosemantic lexical occurrences, on the systemic ordering of kinds of complementations (reflected by the underlying order of the items included in the focus), on definiteness, on lexical semantic properties of words and (if spoken input is used) on the position of the **intonation center** (sentence stress). The primary position of the intonation center is at the end of the sentence (where it need not be phonetically realized by a specific stress), but also in another (secondary) position the intonation center marks the most dynamic part of the sentence (focus proper), cf. (2), where the underlying order is as indicated by (2'):

(2) Charles has found your PEN in a box lying on the table.
(2') (Charles) (box ((Rel) (table) lie)) find ((you) pen)

After several years of research in this domain, which has included psycholinguistic experiments with Czech and German sentences, as well as investigations with native speakers of English, we are convinced that in the individual languages there exists a basic ordering of the kinds of complementations of every verb (noun, adjective). We assume that this ordering, called **systemic ordering,** directly determines the underlying word order in the focus, so that if a sentence part A follows another one, B, under systemic ordering, then B is less dynamic than A (i.e. B precedes A in the underlying word order) only if B belongs to the topic. In the topic part of the sentence the underlying word order often differs from systemic ordering. The systemic ordering of some of the main kinds of complementations in English has the following shape: Time - Actor - Addressee - Objective - Origin - Effect - Manner - Directional(from) - Means - Directional(to) - Locative

2. An automatic identification of topic, focus and the degrees of communicative dynamism, discussed in a preliminary way by Hajičova and Sgall (1985), can be based on the following considerations: In languages with a high degree of "free" word order (as in most Slavonic languages), a secondary position of the intonation center is frequent only in spoken dialogues. In technical texts (spoken or written) there is a strong tendency to arrange the words so that the intonation center falls on the last word of the sentence (where it need not be phonetically manifested), of course with the exception of enclitic words.

A general procedure for determining the topic-focus articulation in such languages can then be formulated as follows:

(i) All complementations (participants and adverbials, or arguments and adjuncts) preceding the verb are contextually bound. As for the complementations following the verb, a "main rule" may be stated: the boundary between topic (to the left) and focus (to the right) can be drawn between any two elements, provided that those belonging to the focus are arranged in the surface word order in accordance with systemic ordering of the kinds of complementations.

(ii) The verb is ambiguous as for its position in the topic or in the focus.

(iii) If a spoken utterance (with its intonation center identified) is analyzed, then similar regularities hold for sentences with normal intonation (intonation center at the end). However, if a non-final element carries the intonation center, then all the complementations standing after this element are contextually bound; for the rest of the sentence, (i) and (ii) hold; the bearer of the intonation center belongs to the focus.

In English the surface word order is determined by grammatical rules to a large

extent, so that intonation plays a more decisive role than in the Slavonic languages. The written shape of the sentence does not suffice here to determine the topic-focus articulation to such a degree as e.g. in Czech. The "main rule" also applies, but otherwise only certain important regularities can be stated here on the basis of word order and grammatical values (especially the articles and other determiners).

In order to be able to reduce the ambiguity of the written shape of the English sentence as much as possible, it is also necessary to take into account certain semantic clues: especially with Locative and the Temporal modifications, it is important to distinguish between specific information (e.g. *on a nice September day, on October 22, 1991, seven months ago*) and items containing just a general setting (e.g. *always*) or being directly (as indexicals) determined from the utterance (*here, today, this year*). The latter examples usually belong to the topic, the former ones typically occurring in the focus. As for the verb, it is important to have access to the verb of the preceding utterance: if the main verb of sentence n has the same meaning as (or a meaning included in) that of sentence n-1, then it belongs to the topic; also verbs with very general lexical meanings (such as *be, have, happen, carry out, become*) may be handled as belonging to the topic. Otherwise (i.e. in the unmarked case), the verb generally belongs to the focus.

3. In the output of the algorithmic procedure completing the parsing of a written English sentence, many ambiguities remain, but it is known that sentences (even in their spoken shape) often are ambiguous as for their topic-focus articulation, so that it should be understood as a good result if the procedure identifies such an ambiguity. The algorithm has been formulated as follows:

(a) The input to our part of the parser is assumed to have passed through the preceding parts, by which the dependency structure of the sentence has been identified, so that also the underlying dependency relations (valency positions) of the complementations (to the governing verb) are known.

(b) If the verb occupies the rightmost position in the sentence and its subject is

(ba) definite (including noun groups with *this, one of the*, etc.), then the verb belongs to the focus getting the index f, and its subject belongs to the topic, which we denote by the index t;

(bb) indefinite, then the subject is (indexed by) f and the verb is t. In either case, the other complementations are handled according to (cb) below.

(c) If the verb does not occupy the rightmost position, then:

(ca) the verb itself is understood as t, if it has a very general lexical meaning (see above), or as f if its meaning is very specific, or else the verb is characterized as intermediate, i.e. ambiguous, abbreviated as (t/f);

(cb) the complementations preceding the verb are denoted as t, with the exception of an indefinite subject and of a specific (i.e. neither general nor highly indexical, see above) Temporal complementation; either of the latter two is characterized as t/f;

(cc) to the right of the verb,

(i) if there is a single complementation, and this is a personal pronoun or another definite noun group, then it is t or t/f, respectively;

(ii) if the rightmost complementation is Temp or Loc, then if it is specific, it is f and otherwise it is t; if it is another kind of complementation, then if it is indefinite, it is f and if definite, it is t/f;

(iii) if there is such an ordered pair A,B to the right of the verb that fails to follow systemic ordering (see Section 2 and the "main rule" above), and B has not been assigned the index t according to (ii), then, for the rightmost such pair, A belongs to the topic (t), and so do all the complementations between A and the verb; the rightmost complementation

of the whole sentence is f (only a personal pronoun following another one is t/f in this position), all those standing between A and the rightmost one are t/f;

(iv) if (iii) does not apply then all remaining complementations to the right of the verb are t/f.

(d) If all the complementations have been determined as t, then

(da) if the verb was t/f after point (ca) and the rightmost complementation is a definite noun group, an indexical word or pronoun, then this rightmost element gets f (this result is abbreviated as t(f));

(db) if (da) does not apply, then both the rightmost element of the sentence and its verb get t/f.

(e) The remaining representations containing no f are discarded.

(f) The complementations with the index t are shifted to the left of the verb, those with f, to the right of it.

Let us add that our algorithm only determines the appurtenance of an element to the topic or to the focus, but does not specify the underlying word order within topic. When implemented (together with a simplified parser), the algorithm was checked with a set of sentences, and it yielded the expected results, cf. the following examples (the notation of which is simplified in that the indices characterizing the underlying structure (cf. (1') above) are left out). NOTE: Our examples concern written English sentences. In its present form, the algorithm handles only the verb and the parts of sentence immediately depending on it; deeper embedded items (esp. adjuncts of nouns) are left aside for the time being.

Examples:

(A) Charles found the pen in a box.

The steps of the analysis (mostly in a simplified notation, without the grammatical indices):

after the application of

(a): $(\text{Charles})_{\text{Act}}$ find.Pret $(\text{pen.Indef})_{\text{Obj}}$ $(\text{box})_{\text{Loc.in}}$
(ca): Charles find.t/f pen box
(cb): Charles.t find.t/f pen box
(cc)(ii) Charles.t find.t/f pen box.f
(iv) Charles.t find.t/f pen.t/f box.f
(f) and resolution of the abbreviation t/f:
Charles.t find.f pen.f box.f (e.g. answering: *Why are the children so happy?*)
Charles.t pen.t find.f box.f (e.g. answering: *How did Charles get the pen?*)
Charles.t find.t pen.f box.f (e.g. answering: *What did Charles find where?*)
Charles.t pen.t find.t box.f (e.g. answering: *Where did Charles find the pen?*)

(B) A Frenchman proved the theorem.
(a) $(\text{Frenchman.Indef})_{\text{Act}}$ prove $(\text{theorem})_{\text{Obj}}$
(ca) Frenchman prove.t/f theorem
(cb) Frenchman.t prove.t/f theorem
(cc)(i) Frenchman.t/f prove.t/f theorem.t/f
(e),(f) prove.f Frenchman.f theorem.f (without topic)
Frenchman.t prove.f theorem.f (e.g. answering: *What did Frenchmen achieve in this field?*)
prove.t Frenchman.f theorem.f
Frenchman.t prove.t theorem.f
theorem.t prove.f Frenchman.f (i.e. pronounced *A Frenchman PROVED the theorem*)
Frenchman.t theorem.t prove.f (ditto)
theorem.t prove.t Frenchman.f (e.g. answering: *Who proved the theorem?*)

(C) At noon Mike awoke.
(a) $(\text{noon})_{\text{Temp}}$ $(\text{Mike})_{\text{Act}}$ awake
(ba) noon Mike.t awake.f
(cb) noon.t/f Mike.t awake.f

(e),(f) Mike.t awake.f noon.f
Mike.t noon.t awake.f

(D) Yesterday we arrived to Nice from Grenoble.
(a) $(\text{yesterday})_{\text{Temp}}$ $(\text{we})_{\text{Act}}$ arrive $(\text{Nice})_{\text{Dir.to}}$ $(\text{Grenoble})_{\text{Dir.from}}$
(ca) yesterday we arrive.t/f Nice Grenoble
(cb) yesterday.t we.t arrive.t/f Nice Grenoble
(cc)(ii) yesterday.t we.t arrive.t/f Nice Grenoble.t/f
(cc)(iii) yesterday.t we.t arrive.t/f Nice.t Grenoble.t/f
(e),(f) yesterday.t we.t Nice.t arrive.f Grenoble.f
yesterday.t we.t Nice.t arrive.t Grenoble.f
yesterday.t we.t Nice.t Grenoble.t arrive.f

(E) Bob met her.
(a) $(\text{yesterday})_{\text{Temp}}$ $(\text{Bob})_{\text{Act}}$ meet $(\text{she})_{\text{Obj}}$
(ca) yesterday Bob meet.t/f she
(cb) yesterday.t Bob.t meet.t/f she
(cc)(i) yesterday.t Bob.t meet.t/f she.t
(d) yesterday.t Bob.t meet.t/f she.t(f)
(e),(f) yesterday.t Bob.t she.t meet.f (i.e. *Yesterday Bob MET her*)
yesterday.t Bob.t meet.t she.f (i.e. *Yesterday Bob met HER (rather than HIM)* or similarly)

References

[Hajičová and Sgall, 1985] Eva Hajičová and Petr Sgall. Towards an automatic identification of topic and focus. *Proceedings of the 2nd Conference of the European Chapter of the Association for Computational Linguistics*, Geneva, 263-267, 1985.

[Sgall, 1986] Petr Sgall, Eva Hajičová and Jarmila Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Ed. by J. Mey. Dordrecht:Reidel - Prague:Academia, 1986.