STOCK OF SHARED KNOWLEDGE - - A TOOL FOR SOLVING PRONOMINAL ANAPHORA

EVA HAJIČOVÁ, VLADISLAV KUBOŇ and PETR KUBOŇ

UFAL
Charles University
Malostranské nám. 25
CS-118 00 Prague
Czechoslovakia

School of Computing Science
Faculty of Applied Sciences
Simon Fraser University
Burnaby, B.C. V5A 1S6
Canada

ABSTRACT

The paper develops further the idea of using the notion of the stock of shared knowledge (SSK) for anaphora resolution following a more subtle treatment of the influence of the topic/focus articulation of the sentence on the degrees of salience of items of the SSK. An algorithmic evaluation procedure of the SSK is formulated taking into account the notions of contextual boundness, syntactic associations, complexity of the sentences and existence/nonexistence of possible competitors, and a general evaluating function is proposed, essential for the process of anaphora resolution. In the present paper the analysis is performed for Czech; however, the considerations are claimed to be of a universal validity, the actual relations between different factors and the values, of course, being language-dependent.

1.INTRODUCTION

In our paper at Coling'90 we followed up the investigations presented in Hajičová, Vrbová (1982) and proposed an algorithm for solving pronominal anaphora with the use of "stock of shared knowledge" (SSK) - an abstract representation of the hierarchy of salience of the items of the knowledge assumed by the speaker to be shared by him and the hearer. The changes of degrees of salience were dependent solely on the bipolar division of the sentence into its topic and focus parts, respectively. In particular, the rules for computing the degrees of salience were specified as follows:

(i) the items referred to in the focus part of the utterance be it by a noun or by a stressed pronoun receive the highest degree of salience (MAX);

(ii) the items referred to by a noun in the topic part of the sentence are activated one degree less (MAX-1) than the items referred to in the focus part;

(iii) a pronominal reference to an item in the topic part of the utterance keeps the activation unchanged;

(iv) the items not mentioned in the given utterance subtract two degrees from their previous activation.

As we stated already in the above mentioned paper, this was only a tentative solution on the way to a more sophisticated approach to organizing "common knowledge" of the speaker and the hearer.

Supported by a thorough linguistic analysis of a large amount of Czech prosaic texts Hoskovec (1989), two possible improvements to the procedure have been suggested Hajičová, Hoskovec, Sgall (in press), namely:

(i) to replace the binary account of topic/focus articulation of the sentence by a more atomic distinction between the contextually bound and non-bound elements of the sentence, thus enriching the numerical system of possible degrees of salience;

(ii) to account explicitly for the empirical observations that items mentioned throughout the discourse are more likely to be referred to than those mentioned only once.

In this paper we would like to argue that other important features should be taken into account in building the new evaluation system for the SSK. We believe that for a more sophisticated treatment of pronominal anaphora, an account of SSK must also allow for:

(iii) a reflection of the topology of the surface structure of the text, in the simpliest form in terms of the distance of the possible antecedent and a

refering expression measured by the number of interfering objects between them with respect to the sentence and paragraph boundaries;

(iv) a capturing of some associations between lexical units describing objects in the text. We have limited our attention only to syntactic associations between governing and dependent words in the syntactic structure of the sentence. More general treatment of associations requires the use of semantic and/or pragmatic information (eg. semantic features, knowledge base etc.) which is not taken into consideration in the present version of the algorithm, but forms a promising subject of further investigations of possible improvements of the algorithm.

Taking these observations into account, we present a new, enriched model of SSK here. In Section 2 we briefly discuss the relevance of the above mentioned features for anaphora resolution. Section 3 gives a proposal of the organization of SSK, together with the rules evaluating changes in degrees of salience of its items and a general algorithm for reference assignment based on the use of SSK. The possibility of customizing the algorithm for the purposes of a special language under consideration (in our case Czech) is discussed in Section 4.

2. MOTIVATION

In our analysis, we work within the framework of the functional generative description (see Sgall, Hajičová and Panevová, 1986). We represent the meaning structure of a sentence as a dependency tree rooted in the main verb, the nodes of the tree being labelled by lexical and morphological meanings. The edges denote the underlying grammatical relations between nodes. All nodes of the tree can be either contextually bound (CB) - if the objects they denote are "given", "known" from the context - or non-bound (NB) - if they introduce new information into discourse.

The meaning of a sentence represented by such a tree is then viewed as divided into two parts - a topic (T), "stating" what the sentence is about, and a focus (F), commenting or developing the topic.

The topic-focus articulation (TFA) of a sentence can be specified according to the sentence structure as follows (cf. Sgall 1979):

(i) F contains the main verb iff the verb is NB;

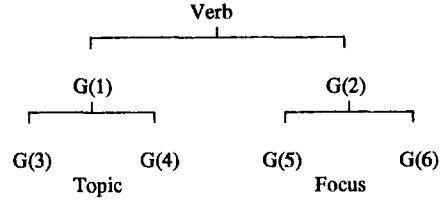
(ii) F contains all daughter nodes of the verb which are NB, together with all nodes subordinated to them (which in turn are either NB or CB);

(iii) if the verb together with all daughter nodes is CB (and, therefore, none of (i),(ii) applies), F is defined with respect to a deeper embedded node. (This case is rather rare and we do not consider it in our analysis for the sake of simplicity.)

(iv) T consists of all the nodes not contained in F.

Thus, for the purpose of this paper, only the difference between NB nodes and CB nodes on the first level of dependency is taken into consideration while specifying TFA of a sentence. We would like to show in the sequel that there is a linguistic evidence which suggests that deeper levels of syntactic embedding (at least the second level) be accounted for in the resolution of anaphora.

For the sake of simplicity, we represent the sentence schematically:



where G(1) is a group of CB nodes on the first level of dependency (belonging to T),

G(3), G(5) are CB nodes on the second level of dependency (belonging to T and F respectively),

G(2) is a group of NB nodes on the first level of dependency (belonging to F) and

G(4), G(6) are NB nodes on the second level of dependency (belonging to T and F respectively).

2.1 Level of dependency in a syntactic tree

Let us introduce one of the examples which show the necessity of further extension of the scale in the SSK. Consider the following sample of text - Hoskovec (1989):

Ex.1:
(1) At the railway station I saw a *dog* with long *ears*.
(2) It was funny to observe *them* dangling in the wind.
(3) I wondered how *he* happened to get there.

According to our Coling '90 paper there is no distinction in the SSK between *dog* and *ears*. Both are contained in focus of (1), which means that they have the highest degree of salience in the next sentence. Such an account does not explain the fact, that the above introduced order of sentences is

possible and the order (1)-(3)-(2) does not constitute a coherent text - it seems to be impossible to refer to *ears* from the third sentence using the personal pronoun *them* as a refering expression.

The scheme of the syntactic tree as introduced above offers us a key to the solution of this problem. From this point of view there is a distinction between *dog* and *ears* in the sentence (1). According to our scheme, the word *dog* stands in the position G(2), the word *ears* is in the position G(6). Both are contextually non-bound.

Thus, examples along this line seem to suggest that the modified SSK has to take into account the distinction between immediate members of a respective verb frame and words which are embedded on a deeper level of the syntactic tree.

2.2 Contextual boundness and non-boundness

The distinction between contextually bound and non-bound elements is also significant. Let us consider the following example from Hoskovec (1989):

Ex.2:
(4) At the railway station I saw *their dog*.
(5) I realized *they* would look for *him* the whole afternoon.
(6) I wondered how *he* happened to get there.

Although this sample text seems to have the same distribution of pronouns as (1)-(3), the difference between the two texts shows when we change the order of sentences to (4)-(6)-(5). In the latter case, the change of the order is possible . Since the sentences (1) and (4) differ only in contextual non-boundness of *long ears* vs. contextual boundness of *their*, respectively, both expressions being on the second level of dependency, we conclude that the distinction between contextual boundness/nonboundness of the nodes in the syntactic tree of the sentence is important for the resolution of anaphora and, therefore, must be captured by the new version of SSK.

2.3 Syntactic associations

The notion of syntactic associations is introduced by means of slightly modified examples found in technical texts. Let us start with the following sample text:

Ex.3:
(7) In the residence quarter of Brno it is possible to find a villa of *professor Schmidt*.
(8) It was built during the thirties.
(9) *His* other two houses are to be found in Olomouc and Jihlava.

In this case the assignment of *him* to its antecedent is straightforward; although the expression *professor Schmidt* is in the focus part of (7), it does not depend directly on the governing verb and, moreover, it is contextually non-bound. At the first sight this seems to be a counterexample to the above introduced scheme of the role of CB and NB elements of a sentence, namely, to the impossibility of referring to NB-nodes on the second level of dependency by means of personal pronouns across one embedded sentence (see Ex.1).

However, we believe that the difference between (1)-(2)-(3) and (7)-(8)-(9) lies in the fact, that *his* is in the third sentence accompanied by the full noun reference to the villa using a similar word (house), which certainly influences the salience of the item *professor Schmidt*. The structure of a noun phrase governed by villa in (7) is the same as the dependency structure of the noun phrase governed by houses in (9), therefore also the salience of the item *professor Schmidt* is evidently higher than without that association. We can support our observations with the modified example:

Ex.4:
(7) In the residence quarter of Brno it is possible to find a villa of *professor Schmidt*.
(8) It was built during the thirties.
(9a) *He* was known as a collector of paintings of young local painters.

In our opinion, the process of assigning the antecedent *professor Schmidt* to the refering expression *him* is not as straightforward as in Ex.3; indeed, some of the hearers have difficulties with accepting Ex.4 as a valid text.

The degree of the influence of syntactic associations on anaphora resolution can vary for different languages. It is also clear that at least a small stock of related notions plays a very important role in this mechanism. We will discuss these problems more in detail in the Sect. 4 of this paper, where we show the approach for a particular language under consideration (Czech).

2.4 Topology

We can use Ex.3 to show another important fact which has an influence on the reference assignment. The sentence (8) is a very simple one, in particular, it does not introduce any new element into the SSK except the word *thirties*. The situation is very different, if we replace (8) by (8a):

Ex.5:
(7) In the residence quarter of Brno it is possible to find a villa of *professor Schmidt*.
(8a) The building was built by a group of architects in late thirties.
(9) *His* other two houses are to be found in Olomouc and Jihlava.

The reference by *him* in (9) is in this case still possible, but the text is not as clear as in Ex.3. Any other new element in (8a) makes the reference almost unclear.

Supported by this observation, we believe that also the linear distance between an antecedent and a refering expression influences to some extent the salience of the referred item.

It is clear that the function which expresses the degree of salience is not continuous. The end of the paragraph seems to have a strong effect: it leads to a drop of the salience of almost all possible antecedents except for those the activation of which has been established by repeated mentioning in the previous paragraph. The exact values of the function are now the objects of intensive investigation. We discuss some results of our investigations into this problem in Sect. 4 below.

2.5 Existence of competitors

The last feature which is considered in our system is the role of competing elements. We can demonstrate the problem by means of a slight change of (8a), which introduces a new competing element into the text:

Ex.6:
(7) In the residence quarter of Brno it is possible to find a villa of *professor Schmidt*.
(8b) The building was built by *architect Hovorka* in late thirties.
(9) *His* other two houses are to be found in Olomouc and Jihlava.

In this case *professor Schmidt* is no longer available as an antecedent for pronominal anaphora since *architect Hovorka* has a greater degree of salience and the same morphological categories.

All previous examples show the necessity of including into the evaluation procedure of the SSK not only the notions of contextual boundness, but also associations, complexity of the sentences and existence/nonexistence of possible competitors. Their role in the evaluation procedure is described more in detail in the following paragraph.

3. THE GENERAL EVALUATING PROCEDURE

Before we start the explanation of our evaluation procedure, we must make clear that we restrict ourselves in our considerations to those items of knowledge (i.e.the mental representations of the objects of the outer world), referred to in the sentence by noun or by a pronoun. The starting conditions for the evaluating procedure are then as follows:

We assume that our procedure is a part of a larger complex system which is able to provide our procedure with the result of syntactico-semantic parsing of any sentence in the form of a dependency tree as a representation of the meaning of the sentence in the sense of Sgall, Hajičová, Panevová, (1986). We do not assume the existence of any special knowledge base, any semantic evaluation procedure or semantic features present in the syntactic tree. For the time beeing we restrict ourselves to those items (mental objects) that are rendered by nouns or pronouns.

The SSK as a basic data structure can be viewed in our modified account as a set of items, which represent all mental objects rendered by nouns or pronouns from the respective text. Each data entry has the form of an ordered quantuple:

<LEX,MORPH,LAST,SYNT,OCCUR>,
where

LEX
represents the lexical value of the item;

MORPH
is a set of morphological characteristics of the word (e.g. gender, number, etc.). These characteristics are used in so-called morphological filter, which filters out the impossible antecedents of the referring expression.

LAST
are the coordinates of the latest occurrence of the word or of the pronominal reference to it. These coordinates are composed of the "surface" and "deep" part. The "surface" coordinates contain the number of the sentence and a serial number of the node in the sentence structure and they serve as a basis for the "topological" part of the evaluation procedure.

The "deep" part contains the code for the position of the word in the syntactic tree as introduced above (G(i)). This information determines the contextual (non)boundness of the word.

SYNT

contains the data about the syntactic structure of the sentence where the respective LEX was mentioned for the last time. The structure is represented only partially, by means of pointers, which point to the governing node and also to all dependent nodes if they are contained in the SSK. This system of syntactic pointers serves as a basic data structure for the simple handling of associations.

OCCUR

is a pair of integers which represent the number of occurrences of the given item both from the beginning of the text and from the beginning of the paragraph.

The algorithm processes the given text sentence by sentence. It receives the dependency tree of a new sentence from the syntactico-semantical preprocessor, together with the list of all the pronominal referring expressions contained in the sentence. Each referring expression in this list carries the information about its position in the sentence (the same as LAST) and about its form (weak or strong pronoun, etc.). Using SSK, the algorithm finds the antecedents for all referring expressions. Afterwards, it changes the degrees of salience of the items in the SSK and reads the next sentence from the input.

Having stated the general idea of the algorithm, we can describe the evaluation process in more detail as follows:

3.1 Algorithm:

(i) Read an input (the syntactic structure of the new sentence and the list of referring expressions). For every referring expression R_i, $i=1,..,k$ in the list do the following (preserve the order of the referring expressions with regard to hierarchy of communicative dynamism in case that the sentence contains more than one referring expression):

a) Use the morphemic filter to filter out all units from the SSK which cannot be considered as possible antecedents of the refering expression R_i.

b) Apply the evaluating function E(w) to all possible antecedents $W_i^1, W_i^2, \ldots, W_i^k$ and sort them according to the obtained results from the most probable antecedent W_i^1 to the least probable antecedent W_i^m.

(ii) For all referring expressions R_i and all results of evaluation W_i^j, $i=1,..,k$; $j=1,..,l_i$ find the best solution. Thus we are looking for the optimal solution of anaphora for the sentence as a whole, since some "best" solution for the particular expression can block successful reference assignment for other refering expressions (Cf. examples in Hajičová, Kuboň, Kuboň, 1990).

Generally, this is a computationally expensive solution but in practice the number of refering expression and possible antecedents is strongly limited and, therefore, this phase does not impose a serious restriction on the performance of the algorithm.

(iii) Update the data in the SSK
- change OCCUR if the item was mentioned or reffered to in the current sentence
- add items mentioned for the first time into the SSK
- remove all the items with degrees of salience function smaller than some constant THRESHOLD (which may vary with respect to the type of the text and the particular language).

The function of salience has the form:

$$S(w)=O/(N*(N-L+1)),$$

where w is the item of SSK under consideration,
O is the number of occurrences of the item in the given paragraph
N is the serial number of the current utterance (in the given paragraph)
L is the serial number of the utterance in the paragraph where this item was mentioned for the last time

3.2 The general evaluating function

This function is essential for the whole process of anaphora resolution. Also, it is considerably more dependent on the language under consideration than all the other parts of the process. For this reason we have divided its description into two parts. In this section we describe the function only generally. The method of customizing all the constants according to the needs of a particular language (in our case Czech) is described in Sect. 4 below.

The basic form of the function is:

$$E(w) = \sum_{i=1}^{n} (c_i * f_i) ,$$

where f_i is a function describing the value of the factor$_i$

c_i is a constant expressing the weight of the factor$_i$

4. THE METHOD OF THE CUSTOMIZATION OF THE EVALUATING FUNCTION

In this paragraph we want to show the method chosen for finding the values of the c_i and f_i for the particular language.

4.1 First step of the method is to find the form of f_i for all factors taken into account. All functions should have a common value range. The balance of influence of all factors is achieved by the help of constants c_i. After a complex examination of Czech texts (with a special stress on technical texts) we have come to the following results:

a) Contextual boundness - the word *w* is either bound or nonbound, therefore

$f_1(w) = 100 <=>$ w is contextually bound

$f_1(w) = 0 <=>$ w is contextually nonbound

b) Underlying structure - for the definition of this function it is necessary to extend our schema from the paragraph 2 deeper than to the second level of dependency. The rule for the extension is the following:

All deeper levels consist only of nodes belonging to groups G(3-6) so that any governing node in the topic governs nodes G(3) and G(4), the governing nodes from focus govern nodes G(5) and G(6).

The function f_2 has been assigned the following tentative forms:

$f_2(w) = 70$ for w in a position of G(1)
$f_2(w) = 100$ for w in a position of G(2)
$f_2(w) = 50$ for w in a position of G(3)
$f_2(w) = 0$ for w in a position of G(4)
$f_2(w) = 50$ for w in a position of G(5)
$f_2(w) = 30$ for w in a position of G(6)

The motivation for this distribution of values can be found in Hoskovec (1989).

c) Associations - if the word w_1 depends directly on the word w_2, it shares a part of the value of $E(w_2)$. We do not restrict the dependency only to the immediate dominance, but the words on a deeper level share less of the value $E(w_2)$. We also take into account that one word can be in principle associated with more than one other member of the SSK. Therefore the form of the function f_3 is the following:

$w_1, .., w_n$ are the governing words of w so that w_i are ordered according to the syntactic level (w_n is the immediate governor of w)

$$f_3(w) = \sum_{i=1}^{n} (1/2^i) * E(w_i)$$

d) Linear distance - this function is quite simple, it is only necessary to count the linear distance of w and the possible refering expression. The counting is easy - we count only the members of SSK.

The function is simple:

$$f_4(w) = 100/((\ln d)+d)$$

where d is a distance between the word w and a possible refering expression.

4.2 There is of course a significant difference between the way of computing f_i and c_i. The latter is a constant, which describes the role of particular factors in the respective language.

For the evaluation of weights c_i we use the following method:

In real texts we look for pieces of text with complicated referring structure. Any such text is modified by adding or removing items. The results are given to a group of randomly chosen native speakers, who should mark the understandability of all texts. One example of this method is given here by the modification of sentences (8) and (9) in Ex. 4 and 5 above.

The basic constraint on c_i is described in the following equation:

$$\sum_{i=1}^{n} c_i = 1$$

which means that every c_i describes the role of factor i in percents. This constraint serves for the purpose of keeping the balance between particular factors under control. It is also useful in the case of some future extension of the whole system by adding new factors.

There can of course be any other constraints according to the needs of a particular language. We do not have any additional constraint for Czech in the moment.

The work on collecting material for the tests on c_i is now in progress. The following constants were chosen as initial values :

- contextual boundness and non-boundness	0.25
- syntactic structure of the sentence	0.25
- associations	0.25
- linear distance	0.25

5. CONCLUSION

The previous analysis has been done for Czech. We are far from claiming that every language would

reflect the same relations between factors which can help to solve the pronominal anaphora. We have only tried to show how certain factors can result in a more sophisticated treatment of anaphora in NLP.

Our mechanism is designed as an open system, the nature of all functions mentioned here enables to add any number of other phenomena which can help to solve the problem of anaphora resolution.

Our approach is substantially different from the approach of e.g. Alshawi (1987) or Rich, Luper-Foy (1987). Our system does not need any knowledge base except the special thesaurus of related notions. It would be very interesting to combine our approach and the approach of Alshawi in some experimental NLP system.

ACKNOWLEDGEMENT

the work on this paper was carried out under the project of the IBM Academic Initiative.

REFERENCES

Hyian Alshawi (1987): Memory and context for language interpretation, Cambridge University Press.

Eva Hajičová, Petr Kuboň and Vladislav Kuboň (1990): Hierarchy of Salience and Discourse Analysis and Production. In: Proceedings of Coling '90, Helsinki

Eva Hajičová, Tomáš Hoskovec and Petr Sgall (in press), Discourse Modelling Based on Hierarchy of Salience; to appear in Prague Studies in Mathematical Linguistics 11

Eva Hajičová, Jarka Vrbová (1982): On the Role of the Hierarchy of Activation in the Processes of Text Understanding. In: Proceedings of COLING 82, Ed. by J.Horecký, Amsterdam.

Tomáš Hoskovec (1989): An Activation Based Model of Discourse (Towards a negative delimitation of cognitive problems), In Proceedings of the International Workshop The Notion of Cognitive in Linguistics, Sofia, Svyat - Benjamins (to appear). Elaine Rich, Susann Luper-Foy (1987): An Architecture for Anaphora Resolution, MCC Technical Report Number ACA-HI-393-87,

Petr Sgall (1979): Towards a Definition of Focus and Topic, Prague Bull. of Mathematical Linguistics 31,3-25;32,1980,24-32

Petr Sgall, Eva Hajičová, Jarmila Panevová (1986): The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Dordrecht - Prague.