Collecting and Employing Selectional Restrictions

Andreas Wagner
Seminar für Sprachwissenschaft
University of Tübingen, Germany
wagner@sfs.nphil.uni-tuebingen.de

Mattia Mastropietro
Department of Computer Science
Computational Linguistics Group
University of Zurich, Switzerland
mastro@ifi.unizh.ch

Abstract

It is a well-known fact that many natural language sentences may have a large number of possible syntactic analyses with semantically inconsistent interpretations. In this paper, we focus on *selectional restrictions* as a device for lexical and syntactic disambiguation. We will briefly sketch the notion of selectional restrictions from a practical point of view. We provide an example of how such constraints can be used for ambiguity resolution. In section 2 and 3, we address the topic of this article, the problem of acquiring selectional restrictions. We discuss different approaches, in particular methods based on statistical corpus analysis. The role of selectional restrictions in a concrete NLP system is shown in section 4.

1 Introduction

The problem of lexical and syntactic ambiguity is encountered rather often in natural language processing. Theoretical linguistics has made no attempt to specify how a competence model might be implemented or approximated by psychologically plausible processes in language perception and production. Therefore, natural language processing has been crucial to fostering the task of efficiently selecting the most appropriate analysis among the set of possible "candidates". In the fields of psycholinguistics and natural language technology, methods for disambiguation are a widely studied area. In this section we present the basic idea of how *selectional restrictions* ease the computing of an appropriate reading of a sentence.

1.1 Selectional restrictions

Selectional restrictions can be defined as description of semantic constraints between lexical elements. More exactly, a predicate (e.g. a verb) imposes semantic constraints on its arguments (e.g. nominal complements). Such constraints reject semantically anomalous sentences like

Ex. 1 The stone thinks.

The verb `think` selects a subject with the feature `human`, which suggests that words labeled with `inanimate` are rejected.
Selectional restrictions point out the semantic compatibility between a predicate and its arguments: `think` is not compatible with a subject like `stone` that lacks the feature `human`.

1.2 Selectional restrictions and disambiguation

The above example illustrates that selectional restrictions include semantic information by checking the compatibility between a predicate and its arguments. This information can be applied for disambiguation by rejecting analyses with incompatible predicate-argument combinations. Consider the following example (taken from [All94]):

Ex. 2 The dishwasher read the article.

In this example lexical ambiguity is involved. Both the subject noun and the object noun may have at least two different meanings: `dishwasher` can denote either a person or a machine. `article` can even refer either to a text (e.g. in a newspaper) or to a single word (a determiner). Example 2 demonstrates how selectional restrictions can be helpful for resolving the combinatorial explosion of lexical analyses. In the given example, the subject `dishwasher` is *not* ambiguous, it denotes a *person*. This is due to the fact that `read` prefers a subject labeled with an attribute `human`. However, this sentence also illustrates the limitation of this method: Selectional restrictions *cannot do all the work*, i.e. there are ambiguities that selectional restrictions are not able to resolve. In the given example, it is the case with the word sense of `article`, since human intuition allows reading a text and a single word. Thus the semantic constraints that `read` imposes on its object do not provide the information needed for the disambiguation of `article`. (For this task, context information is required.)

1.3 Is a violation of selectional restrictions always semantically anomalous?

Deviations from selectional restrictions need not necessarily result in semantically anomalous expressions. In many cases, such expressions can be interpreted metaphorically, for example

Ex. 3 The car drinks gasoline.

Thus, the term *selectional preferences* describes the phenomenon more appropriately than the traditional term *selectional restrictions*, which may be too "restrictive".

There is another motivation for using the term *preference*: Predicates seem to select different classes with different preference strengths. Consider example 2 again. As we pointed out above, `read` prefers the classes `text` and `word` as possible objects. If a human reader wants to assign an interpretation to example 2 without considering a special context, he is likely to focus on the `text` meaning. So empirical knowledge indicates that `read` prefers the class `text` stronger than the class `word`. The term *preference* covers the idea of gradedness better than the "all-or-nothing term" *restriction*.[1]
Another important issue for the following sections is the matter that a small set of semantic features or semantic class labels is often insufficient for formulating selectional preferences. The semantic constraints that a predicate imposes on its arguments may be idiosyncratic. For example, `diagonalize` selects an object that denotes a matrix. A rather elaborated system of features is required to code such constraints accurately. This makes the task of acquiring selectional restrictions more difficult.

[1] Nevertheless we will use both terms in the remainder of this article.

One of the main scopes of this paper is to show some techniques allowing to approximate a large coverage with selectional restrictions. In section 2, we will motivate the acquisition of selectional restrictions and discuss possible resources for that task. In section 3, we sketch three methods to acquire selectional restrictions by means of statistical corpus analysis. Finally, the role of selectional restrictions in a NLP system is described.

2 Acquisition of selectional restrictions

2.1 Motivation

As we saw in the previous section, the acquisition of selectional restrictions is a non-trivial task. The useful application of selectional preferences for disambiguation has been shown successfully in NLP systems like the *The Core Language Engine* [Als92]. Selectional restrictions have been incorporated into its lexicon in order to be exploited during the disambiguation process. A further motivation for the acquisition of selectional restrictions is providing such information in dictionaries for human users. This is especially useful for foreign language learners, since they have to gain knowledge about idiosyncratic constraints on predicate-argument combination (e.g. that `rise` can select an abstract subject like *temperature*, but *ascend* cannot).

2.2 Possible resources

Two kinds of resources are currently used for the acquisition of selectional restrictions: *human intuition* and *text corpora*.[2]

Gathering selectional restrictions "by hand" is still a common approach. It is particularly useful for a lexicon which is used within an NLP system for a limited domain. Such a lexicon is usually so small (up to 1000 words in current systems) that hand-coding is most appropriate. The builder of such a lexicon can take into account idiosyncratic usages of words (with respect to the concerning domain), which may not be captured by statistical corpus analysis. For large, domain-independent lexicons, manual acquisition of selectional restrictions is time- and labor-intensive, and the danger of getting incomplete and inconsistent results increases considerably. (Nevertheless, there are efforts to build large-scale lexical and world knowledge bases manually, e.g. the CYC project.)

Statistical extensions of linguistic theories have recently gained popularity in the field of natural language processing. It has been widely recognized that pure rule-based methods suffer from a lack of robustness in resolving uncertainty with respect to selecting the most appropriate analysis among a set of possible analyses (in particular if none of these analyses satisfies all constraints imposed by the lexicon and the grammar). If this kind of uncertainty arises, a statistical model can select the analysis with the highest probability. In speech recognition, for example, statistical methods have improved performance significantly.

Psychological investigations indicated that people register frequencies and differences of frequencies. Additionally, people prefer analyses that have been experienced before, which implies that the preference is influenced by the occurrence frequencies of analyses.

[2] For a discussion of the advantages and drawbacks of different resources for the extraction of lexical information in general, cf. [BP96].

This is a justification for quantifying preference by means of probabilistic models (with probabilities estimated according to frequencies observed in the examined corpus).

Concerning our topic, statistical corpus analysis provides the possibility of (semi-)automatically acquiring selectional restrictions on a broad empirical basis. This approach yields consistent information on how words are used. Handling a large number of words automatically, it saves human labor. However, subtle idiosyncratic (perhaps domain-dependent) word uses may not be captured because of lacking statistical evidence. A special advantage with respect to the "preferential" character of selectional restrictions is that preference strength can be captured by probabilistic models.

In summary, manual coding of sectional restrictions (and of lexical information in general) is preferable for building up a relatively small lexicon suitable for a limited domain. Statistical corpus analysis is more appropriate if broad coverage of lexical items is required.

3 Statistical acquisition of selectional restrictions

In this section, we sketch three approaches used for acquiring selectional restrictions by means of statistical corpus analysis. First we introduce the parts of information theory that are relevant for these approaches.[3]

All approaches described here are based on an *ontology* of semantic classes, in terms of which selectional restrictions are represented. Of course it is possible to collect pure information about selectional preferences of individual nouns, but such word-based models do not capture semantic generalizations. Additionally, they lead to practical problems, because they depend more on the peculiarities of the examined corpus than class-based models do. For example, class-based models are able to capture admissible verb-noun combinations that accidentally are not present in the particular corpus.

3.1 Information theoretic fundamentals

Information theory deals with coding information as efficiently as possible. In the framework of this discipline, information is usually coded in bits, and the number of bits is the measure of the amount of information. An example: Suppose we have to code a randomly generated sequence of signs. The most obvious way to do this is to represent each sign by a bit sequence of uniform length. However, if the probabilities of the individual signs (reflected by their frequencies of occurrence) differ significantly, it is more efficient to assign shorter bit sequences to more probable (and thus more frequent) signs and longer bit sequences to less probable (and less frequent) signs. Thus, there is a reciprocal relationship between the probability of an event and the corresponding bit code length, and hence the amount of information the event carries.

The lower bound of the average code length needed to represent a value of a random variable X is called *entropy*, which is given by

$$H(X) = -\sum_x p(x) log\, p(x) = \sum_x p(x) log \frac{1}{p(x)} \qquad (1)$$

with log base 2, due to binary coding.[4]

[3] For a comprehensive introduction to information theory cf. [CT91]

[4] It can be shown that assigning $\lceil log \frac{1}{p(x)} \rceil$ bits to an event x with probability $p(x)$, one can achieve an average code length that is at worst one bit greater than entropy.

Relative entropy is a measure of the distance between two probability distributions p and q. More exactly, it quantifies the cost (with respect to amount of information) of assuming distribution q when the real distribution is p. Relative entropy is defined as

$$D(p\|q) = \sum_x p(x) log \frac{p(x)}{q(x)} = \sum_x p(x)(log \frac{1}{q(x)} - log \frac{1}{p(x)}) \tag{2}$$

If you use q to construct a code for a random variable X with probability distribution p, the average code length is $D(p\|q)$ higher than the entropy of X.

3.2 Pereira and Tishby

To get the noun classes required for gathering selectional restrictions, [PT92] examine the distribution of nouns in a corpus. For each verb they determine the distribution of nouns at the object position. The results can be represented in a matrix like the following:

Nouns	Verbs						
	buy	end	wash	touch	open	start	time
conference	0	1	0	0	1	1	1
closet	1	0	1	1	1	0	0
door	1	0	1	1	1	0	0
meeting	0	1	0	0	1	1	1
mouth	0	0	1	1	1	0	0
number	0	0	0	0	0	0	0
reply	0	1	0	0	1	1	1
store	1	0	1	1	1	0	0
window	1	0	1	1	1	0	0

(This is a simplified representation adopted from [Cha93]. '1' stands for a possible verb-object combination, '0' for the opposite.) By this way, every noun can be assigned a vector (the respective line in the matrix). Then nouns with similar vectors are grouped in classes. (Similarity of vectors is measured by relative entropy.)

Based on these classes, selectional restrictions can be deduced. For example, "open" selects two classes: one containing nouns like "door", "mouth", or "window", and one containing "conference", "reply" etc. This corresponds to two senses of "open": a concrete and an abstract sense. These dates can be used for disambiguation: In the expression "open a door" the verb sense differs from that in "open the conference". However, since the vector of "number" does not fit either class "open" prefers, "open a number" indicates a violation of selectional preferences.

3.3 Resnik and Ribas

In contrast to [PT92], [Res93] uses an already existing large-scale taxonomy of semantic noun classes, namely the noun part of WordNet (cf. [M$^+$90]) to acquire and represent selectional preferences. He quantifies the *selectional preference strength* of a predicate with respect to a certain argument, i.e. how strong the predicate constrains its arguments semantically,[5] by relative entropy. To be more exact, he measures the distance between the probability distribution of noun classes occurring in a certain syntactic

[5]For example, "eat" has a greater selectional preference strength for its object then "see".

relationship to *some verbs* (e.g. as an object) and the probability distribution of noun classes occurring in this relationship to a *particular* verb:

$$D(p(c|v,s)||p(c)) = \sum_{c} p(c|v,s) log \frac{p(c|v,s)}{p(c|s)} \tag{3}$$

(where c is a variable over noun classes, v stands for a verb and s for a syntactic relationship). This measure describes the cost of assuming that the distribution is $p(c|s)$ whereas in reality it is $p(c|v,s)$, and thus "*[...] the cost of not taking the predicate into account. Therefore, in a very direct way, the selectional preference strength of a predicate can be understood as the amount of information that it carries about its argument.[...]*" [Res93, p. 58]

But overall preference strength does not tell *which* classes a verb prefers. The preference of a certain class is quantified by a measure Resnik calls *selectional association*:

$$A(v,s,c) = p(c|v,s) log \frac{p(c|v,s)}{p(c|s)} \tag{4}$$

(Note that this is the addend of the sum in equation (3) that corresponds to class c.) This formula takes into account two things: the *absolute probability* of c occurring as argument s of verb v and the *deviation* of this probability from the general probability of c appearing as argument s.

Resnik's results fit human intuition. For example, they can serve for disambiguating "baseball" in "hit some baseball" vs. "play some baseball": Among the WordNet classes that contain "baseball", the class with the highest selectional association is `object` for "hit" and `game` for "play", respectively.

[Rib94] employs Resnik's selectional association measure to select a subset of noun classes that represent the selectional preferences of a given verb. He uses a simple, heuristic algorithm: In order to exclude noise, only classes with a selectional association value above a certain threshold are considered. Among the remaining set of "candidate classes", the class with the highest selectional association is selected for the final result. This class and all its hyponyms and hyperonyms are removed from the set of candidates to avoid overlapping of classes. This procedure is repeated until the candidate set is empty. The algorithm yields a set of classes that is a non-redundant representation of selectional preferences at an adequate generalization level.

3.4 Li and Abe

Like Resnik and Ribas, Li and Abe gather selectional preferences on the basis of WordNet. However, while Ribas employs a heuristic algorithm to retrieve a representation of selectional preferences that is both compact and accurate, [LA95] apply an information theoretic notion called *Minimum Description Length* (MDL) to achieve that goal.

> [...] MDL is a principle of data compression and statistical estimation from information theory, which states that the best probability model for given data is that which requires the least code length in bits for the encoding of the model itself and the given data observed through it. The former is called the 'model description length' and the latter the 'data description length'.[...]

Li and Abe represent the selectional behaviour of a verb (with respect to a certain argument) as a *tree cut model*, as they call it. Such a model consists of a horizontal cut through the noun hierarchy tree,[6] which yields a set of classes that form a partition of the noun senses covered by the hierarchy, and a probability distribution for this set of classes. Each class is assigned the probability of its occurrence at the given argument position of the given verb.

To illustrate the idea, consider a simple semantic hierarchy with the root ANIMAL, which has two daughter nodes, BIRD and INSECT. The descendants of these nodes are the leaves of the hierarchy, representing simple word senses: swallow, crow, and eagle are the daughters of BIRD, bug and ant are the daughters of INSECT. Possible tree cuts through this little hierarchy are: [ANIMAL], [BIRD, INSECT], [BIRD, bug, ant], [swallow, crow, eagle, INSECT], and [swallow, crow, eagle, bug, ant]. Now suppose there are ten instances of the verb "fly" in the corpus, and the respective subjects are "swallow" (3 instances), "crow" (1 instance), "eagle" (2 instances), and "bug" (4 instances). Thus a possible tree cut model (tree cut plus probability distribution) is, for example, ([BIRD, bug, ant], [0.6, 0.4, 0]).

MDL is used to get the tree cut model with the appropriate generalization level. The model description length depends on the number of noun classes in the model. The data description length for the nouns occurring as the verb argument in question is given by

$$-\sum_{n} log\, p_M(n|v, s) \tag{5}$$

where p_M is a probability distribution determined by the tree cut model M. It assigns a noun n the average probability that a member of the class n occurs as the argument s of the verb v. (According to the tree cut model mentioned above, "swallow", "crow", and "eagle" each get the probability 0.2, "bug" 0.4, and "ant" 0.) p_M represents generalized probabilities corresponding to the classes in M.

If the tree cut is located near the root, then the model description length will be low, because the model contains only few classes. However, the data description length will be high, because code for the data is based on the probability distribution of the classes in the model, not on the real probability distribution of the individual nouns. As we saw in section 3.1, the greater the difference between the supposed distribution and the real one is, the longer the code becomes. And the coarser the classification is, the more the corresponding distribution p_M deviates from the real distribution. If the tree cut is located near the leaves, the reverse is true: the fine-grained classification fits the data well, resulting in a low data description length, but the great amount of classes rises the model description length. Minimizing the sum of these two description lengths yields an adequate "compromise" between compactness (expressing generalizations) and accuracy (fitting the data) of the model.

An example: In [LA95] the results for the object of "eat" are reported. The most probable classes included in the acquired tree cut model are `food`, `life_form`, and `quantum`. This generalization level is in accordance with human intuition.

3.5 Discussion

Although the approaches sketched above work quite well, they are far from being the final solution for the given task. Improvements are possible and have been worked out (for example, cf. [Rib95], [LA96]). A great challenge is the reduction of noise, which biases the results. [Rib94] reports problems concerning noun polysemy: Non-intended

[6] This cut need not be a straight line; its course may touch several levels of the tree.

senses accumulate significantly, leading to the acquisition of wrong classes. [LA95] find a relatively high preference of `artefact` for the object of "eat". Such rather surprising results are often, at least in part, due to idiomatic expressions.

The most notable difference between these approaches is that Resnik and Ribas as well as Li and Abe employ a predefined semantic noun hierarchy, while Pereira and Tishby establish noun classes by examining their corpus. Both alternatives have advantages and disadvantages. Collecting classes by retrieving context vectors and clustering nouns in accordance with the similarity of their vectors makes it difficult to determine the semantic content of the retrieved classes. Meaning cannot be discovered straightforwardly from context vectors. This problem does not arise if one uses a class hierarchy like WordNet, because here the classes are supplied with semantic labels. The drawback of a general hierarchy is that it may be inappropriate for specific domains or sublanguages. [BPV96] state: "*[...] In our research, we analyzed different sublanguages and we found that, while in a given domain there are groups of words that are used almost interchangeably, the same words may have no common supertypes in WordNet.[...]*" [BPV96, p. 121] Such domain-dependent similarities (and dissimilarities) could rather be discovered by corpus-based class acquisition (if the corpora are selected accordingly).

4 Employment of selectional restrictions

The Computational Linguistics Group at the University of Zurich is in the process of building an NLP system which will use selectional restrictions. The system is called the *University-Information-System* (UIS). It focuses on answering questions concerning University specific issues over a WWW interface. The query language is German.

The computation of the answer is achieved in several steps. First, the user is prompted to ask a question to the system. The natural language input is then morphologically and syntactically analyzed. The morphological part is taken over by GERTWOL, a morphological analysis program distributed by Lingsoft Corp. GERTWOL provides inflection information for about 100,000 stems and a powerful derivation and composition component. The program is restricted to morphological information. It does not provide valency information or selectional restrictions. To complement the GERTWOL lexicon, the UIS lexicon was built up with lexical information taken from the CELEX database [7] mainly subcategorisation information for verbs. The verb forms in CELEX are classified as auxiliary verb, copula, impersonal verb, reflexive verb and full verb. For each verb one or several subcategorisation classes are provided, defined for example as *accusative object*, *prepositional complement* or *adverbial complements*. The adverbial complements are further divided into eight semantic classes: *locative, temporal, manner, causative, purpose, instrumental, comitative, role.* But CELEX does not provide any semantic information for verbs.

Therefore, we were looking for a more powerful lexicon, which would describe the selectional restrictions of German verbs. We found the Griesbach/Uhlig-Lexicon. It describes the usage of German verbs with example sentences. It contains 3,000 strong German verbs and 6,000 weak verbs. The lexicon is available in machine readable form. We have planed to extract the semantic information of each verb which is classified as *person*, *thing* (German: Sache), *notion* (German: Begriff), and *fact* (German: Sachver-

[7] CELEX was developed at the Max Planck Institute for Psycholinguistics in Nijmegen. It contains lexical information for German, English and Dutch.

halt). This is a rather poor taxonomy but since it is available for a huge number of verbs it will still be of some help if we can find the corresponding classification for nouns. This is not included in the lexicon. We may end up doing a labor intensive hand coding of the relevant nouns if there is no such information available for German.

After finishing the morphological check, the syntactic analysis is done by an ID/LP parser [8]. It was extended to process traditional phrase structure rules as well, in order to increase the flexibility in writing grammar rules. The algorithm is a bottom-up chart parser written in SICStus Prolog.

The parser tries to apply a grammar rule by finding the elements on its right hand side and in case of success inserting an edge into the chart. The goal is to find an edge over the whole sentence. Some ambiguity shall be resolved by the selectional restrictions technique. The verb's selectional preferences will be unified with the complements' features. For this purpose, we have to define a word class hierarchy specifying the relationship between the word groups, as [Als92] proposed in the CLE system. A semantic network for German is not available yet, so we will concentrate on the specific university domain.

Afterwards the appropriate reading is used to generate a logical formula. As a next step, a theorem prover over the logical formulae localizes the answer. The answer will afterwards be extracted from the database and generated via syntactic rules.

References

[All94] James Allen. *Natural Language Understanding*, The Benjamins/Cummings Publishing Company, Inc., second edition, 1994.

[Als92] Hian Alshawi. *The Core Language Engine*, The MIT Press, London, 1992.

[BP96] Branimir Boguraev and James Pusteyovsky. Issues in Text-based Lexicon Acquisition. In Branimir Boguraev and James Pusteyovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 3–17. MIT Press, Cambridge, Massachusetts, 1996.

[BPV96] Roberto Basili, Maria-Teresa Pazienza, and Paola Velardi. A Context Driven Conceptual Clustering Method for Verb Classification. In Branimir Boguraev and James Pusteyovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 117–142. MIT Press, Cambridge, Massachusetts, 1996.

[Cha93] Eugene Charniak. *Statistical Language Learning.* MIT Press, Cambridge, Massachusetts, 1993.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

[LA95] Hang Li and Naoki Abe. Generalizing Case Frames Using a Thesaurus and the MDL Principle. In *Proc. of Int. Conf. on Recent Advances in NLP*, 1995.

[LA96] Hang Li and Naoki Abe. Learning Word Association Norms Using Tree Cut Pair Model. In *Proc. of 13th Int. Conf. on Machine Learning*, 1996.

[8] ID stands for *immediate dominance* and resembles the phrase structure notation, but without specifying the constituent order. LP means *linear precedence* and specifies the linear order of the sister constituents.

[M+90] George A. Miller et al. Five papers on WordNet. CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.

[PT92] Fernando Pereira and Naftali Tishby. Distributional Similarity, Phase Transition and Hierarchical Clustering. In *Proc. of AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 1992.

[Res93] Philip Stuart Resnik. *Selection and Information: A Classed-Based Approach to Lexical Relationships*, Dissertation, University of Pennsylvania, 1993.

[Rib94] Francesc Ribas. An experiment on learning appropriate selectional restrictions from a parsed corpus. In *Proc. of COLING*, Kyoto, 1994.

[Rib95] Francesc Ribas. On learning more appropriate selectional restrictions, 1995. Natural Language E-print Archive: cmp-lg/9502009.