Selectional Preferences from a Latin Treebank

Barbara McGillivray
University of Pisa
E-mail: b.mcgillivray@ling.unipi.it

Abstract

We present a system for automatically acquiring selectional preferences for Latin verbs. We use the *Index Thomisticus* Treebank Valency Lexicon and an enriched version of Latin WordNet as the reference conceptual hierarchy.

1 Introduction

The linguistic community today can rely on large annotated corpora, lexical resources and Natural Language Processing (NLP) tools for several modern languages. Compared with these, Latin is a low resource language. In the past years various projects have started aiming at filling this lacuna. Three treebank projects are ongoing, sharing annotation style: Latin Dependency Treebank (LDT), *Index Thomisticus* Treebank (IT-TB) and PROIEL Treebank.[1] Among the lexical databases, Latin WordNet (LWN) [6] consists of around 10,000 Latin lemmas mapped into the parallel structure of MultiWordNet [3]. The syntactic content of treebanks joined with the semantic information from LWN allows for further research in distributional computational semantics and in several NLP applications: word sense disambiguation, anaphora resolution, parsing, etc.

This paper deals with the issue of automatically extracting semantic information from syntactically annotated Latin corpora. In particular, we aim at extracting the verbs' selectional preferences (SPs), i. e. the semantic preferences of verbs on their arguments. For example, the subject position of the transitive verb *think* is usually filled by lexical items whose semantic properties include being human. The background for this work is the IT-TB Valency Lexicon [5], which collects syntactic arguments of verbs occurring in the IT-TB and represents them in a structured, easily searchable way. It is automatically extracted from this treebank and updated as the annotation proceeds. Following the method illustrated in [1] developed for a cognitive model, we exploit the semantic hierarchy of LWN to map lexical items into concepts and extract their semantic relations among them. Then, SPs are calculated as probability distributions over these semantic features. The result is a rich computational resource for the variety of Latin attested in the IT-TB.

[1] The sizes of these treebanks range between 50,000 words (LDT) and 100,000 (PROIEL). See [5] for references.

2 Background and new contribution

Over the last decade, research into automatic acquisition of SPs from corpora led to several systems which rely on supervised and unsupervised methods. Given a set of argument headwords (for example, *doctor*, *child* as subjects of *think*), these approaches perform a generalization step over unseen cases (*human being*). The generalization problem is represented in WordNet (WN) approaches in terms of preference probabilities over a noun hierarchy, and the goal is to find the appropriate noun classes for each verb-argument pair (the probability of *human being* appearing in the subject position of *think*). This is achieved using statistical tools from information theory, statistical testing and modeling (see [4] for references).

All these models start from single verb-argument occurrences collected from large corpora to infer the probability that the verb's argument position is filled by a class of nouns. Since our dataset is extracted from a small treebank (63,000 tokens), a large number of low frequency verb-noun distributions are observed. Nevertheless, these low frequencies are not necessarily a property of the Latin verbs: they follow from a small which underestimates variance. This can be remedied by grouping the observations into larger clusters. For these reasons, the method illustrated in [1] proved effective when dealing with the peculiarities of our data.

Although we work on the same language and with treebanks comparable in size, our system also differs from the one described in [2]. Bamman and Crane report experiments on extracting SPs from a 3.5 million word Latin corpus which was automatically tagged and parsed using the LDT as training set. The large size allows for better variance estimates and more representative frequencies: this overcomes the problems caused by noisy parsed data. SPs are then extracted through the log likelihood test, a technique also used in collocation extraction. The output consists of association scores between single verbs and single nouns occurring as their arguments. Given the high complexity of the LDT (various authors, genres and time periods), their system makes finer-grained distinctions between the specific usages in different authors, eras, and genres.

While having the same annotation style, IT-TB and LDT differ as to their composition: IT-TB contains works by one author (Thomas Aquinas), belonging into one genre (philosophy). Hence, the variability of the lexicon in the IT-TB is not as extreme as in the LDT, allowing us to treat the corpus as a homogeneous whole. This also implies that a higher number of verb (and noun) instances are typically found that share the same sense (i. e. WN synset). This decreases the probability of finding very low frequency associations between a verb sense and a noun sense, and partially improves the accuracy of the extraction system. Moreover, instead of finding association scores between verbs and nouns in an argument position, we aim at calculating the probability of a WN concept occurring in an argument position for a given verb. This proves to be effective in a lexicographic perspective, where broader semantic classes rather than single words are required.

3 Acquisition of selectional preferences

Alishahi and Stevenson [1] propose a computational model for SP induction in a cognitive framework. In their model, each verb usage is a *frame*, that is a collection of syntactic and semantic features, such as the number of verbal arguments, the syntactic pattern, and the semantic properties of each argument. By semantic properties they refer to the lists of WN hypernym synsets for each word. A *construction* is defined as a collection of frames probabilistically sharing some feature values, for example the transitive construction: a construction clusters frames together based on their syntactic and semantic features. For each (verb, argument position) pair, a probability distribution over a set of semantic properties is calculated; this probability distribution represents the verb's SPs.

We adapted the definitions of frame, syntactic and semantic features to the data at hand. In the IT-TB Valency Lexicon each verbal form occurring in the IT-TB corresponds to a lexical entry recording syntactic, morphological and lexical information on the verb's arguments. For example, the sentence[2]

(1) dominus discipulis formam baptizandi dedit.
Lord-NOM.M.SG disciples-DAT.M.PL form-ACC.F.SG baptize-GERUND-GEN give-PRF.3SG
"the Lord gave to the disciples the form of the baptism."

represents a frame for an active ditransitive occurrence of the verb *do* ("give"); it is recorded in the lexicon as the following subcategorization structure (SCS):[3]

$$do + A\_Sb[nom]\{dominus\}, Obj[acc]\{forma\}, Obj[dat]\{discipulus\} \quad (1)$$

The argument positions (or slots) for the active form of *do* are a nominative subject (A_Sb[nom]), an accusative object (A_Obj[acc]) and a dative object (A_Obj[dat]).[4] The lemmas, or *fillers*, of the lexical items occurring in these positions are: *dominus* ('Lord'), *forma* ('form'), and *discipulus* ('disciple'). We assigned the SCS structures to the set of syntactic features of the frame ($feature_1$).

In order to define the semantic features of each frame, we referred to the LWN database, which contains around 10,000 lemmas aligned with the English WN. The mapping links each Latin synset to an English synset and defines a "lexical gap" when this is not possible. Since the coverage of this resource is low with respect to our data,[5] we semi-automatically added new Latin lemmas to the hierarchy in the following way. For each lemma *L* (for example *abiectio*), we collected its Italian and/or English translations *T* (for example 'avvilimento', 'abbattimento', 'dejection', 'despondency') by using electronic versions of Latin-to-Italian and Latin-to-English dictionaries. Then, we selected the synsets of *T* that are relevant to the

[2] Thomas, *Super Sententiis Petri Lombardi*, IV, Distinctio 8, Quaestio 1, Articulus 3C, Argumentum 2, 3-3.4-1.

[3] 'A' stands for 'Active', 'Sb' for 'subject', 'Obj' for 'object', 'nom' for 'nominative', 'acc' for 'accusative', and 'dat' for 'dative'.

[4] The linear order of these elements in the sentence is not recorded: this choice is due to the relatively free word order in Latin sentences.

[5] 1027 fillers out of 2934 and 90 verbs out of 559 were not present in the lexical database.

senses of L; thanks to the alignment in MultiWordNet, we finally assigned L to the Latin synsets corresponding to these selected senses of T, if any (*humiliatio-humilitas-indignitas*; *contritio*; *demissio*).[6]

For each argument position, the semantic properties of a frame ($feature_2$) are the set of WN hypernyms of the fillers for that slot. For example, the semantic properties for the slot *A_Sb*[*nom*] in (1) are all the hypernyms of *dominus*.

Finally, the semantic properties of the verb belonging to a frame are the list of its WN synsets ($feature_3$): in (1) they coincide with the synsets of *do*.

3.1 Bayesian clustering of frames

In the approach suggested by [1], a frame is clustered into a new construction according to the probabilistic similarity between its features and the features of the frames already included in the construction. This way constructions are created incrementally by means of a Bayesian process. A construction K is chosen for a frame F if it maximizes the probability $P(k|F)$ over all constructions k (including a new construction), that is (after Bayes' theorem) if it maximizes the product $P(k)P(F|k)$. We set the prior probability $P(k)$ to the number of frames contained in k divided by the total number of frames. If we assume that the frame features are independent, $P(F|k)$ is the product of $P_i(feature_i(F)|k)$ for i=1,2,3: $P_i(feature_i(F)|k)$ is the probability that the i^{th} feature displays in k the value it has in F, that is $feature_i(F)$.

$Feature_1$: for the syntactic properties, we estimated $P(feature_1(F)|k)$ by using the following maximum likelihood formula:

$$P(feature_1(F)|k) = \frac{\sum_{h \in k} synt\_score(h,F)}{n_k}$$

where $synt\_score(h,F) = \frac{|SCS(h) \cap SCS(F)|}{|SCS(F)|}$ (*syntactic score*) is the number of syntactic slots shared by h and F over the number of slots in F. This accounts for the degree by which two frames h and F differ in their syntactic patterns (SCSs).[7] For example, let F be the frame given by the verb *coniungo* ('join') + $P\_Obj[dat]\{finis\}, P\_Sb[nom]\{calor\}$ and h be the frame *adiungo* ('join') + $P\_Obj[dat]\{terminus\}$. The algorithm clustered F into a construction containing h: hence, $synt\_score(h,F) = \frac{1}{2}$.

$Feature_2$: for each argument position a in F, $P(feature_2(F)|k)$ is

$$P(feature_2(F)|k) = \frac{\sum_{h \in k} sem\_score_a(h,F)}{n_k} \quad (2)$$

[6] This way, we were able to add 401 new noun lemmas + 90 verb lemmas to 2056 already existing Latin synsets.

[7] Given the high frequency of omitted arguments in Latin sentences, the chances of an exact match between the two SCSs are low. For this reason, we did not define the syntactic score as a binary function.

where $sem\_score_a(h,F) = \frac{|S(h) \cap S(F)|}{|S(h) \cup S(F)|}$ (*semantic score*) accounts for the degree of overlap between the semantic properties $S(h)$ of h and the semantic properties $S(F)$ of F (for argument a). In the previous example, *terminus* ('limit') and *finis* ('end') are the lexical fillers of the $P\_Obj[dat]$ slot for h and F, respectively. The semantic score is $\frac{1}{5}$ because the intersection between the semantic properties of the two words contains one item (*abstraction*), as shown below:[8]

terminus: indefinite_quantity, mensura-modus-quantitas ('measure-quantity-amount-quantum'), abstraction.
finis: locus-punctum ('point'), aetas-circumductio-circumductum-continuatio-periodus-sententia--spatium ('time_period- period-period_of_time-amount_of_time'), abstraction.

$Feature_3$: the probability of displaying in k the value that F has $feature_3$ is

$$P(feature_3(F)|k) = \frac{\sum_{h \in k} syns\_score(h,F)}{n_k} \quad (3)$$

where $syns\_score(h,F) = \frac{|Synsets(verb(h)) \cap Synsets(verb(F))|}{|Synsets(verb(F))|}$ (*synset score*) calculates the degree of overlap between the synsets for the verb in h and the synsets for the verb in F over the number of synsets for the verb in F; n_k is the number of frames in k.

The algorithm uses smoothed versions of all the previous formulas and clusters a frame into a construction after taking into account its syntactic features (relative to the subcategorization pattern) and its semantic properties (relative to both the verb and the lexical fillers of the verb's arguments).

3.2 Selectional preferences

The clustering allows us to perform the generalization step over unseen cases while predicting the probability that a noun n is an argument filler for a verb v in an argument position a; this probability is calculated as the sum of $P_a(n,k|v)$ over all constructions k and can be approximated as the product $P(k,v)P_a(n|k,v)$. $P(k,v)$ is the probability that v occurs in construction k and is estimated as the smoothed relative frequency of v occurring in k. On the other hand, we calculate $P_a(n|k,v)$ as

$$P_a(n|k,v) = \frac{\sum_{h \in k} sem\_score(h,n) \cdot syns\_score(h,v)}{n_{k_v}}$$

where $sem\_score(h,n)$ is the semantic score between the set of semantic properties of the fillers for a in h and the set of semantic properties of n (see formula (2)); $syns\_score(h,v)$ is the synset score between the synsets of v and the synsets of the verb in h; finally n_{k_v} is the number of frames contained in k whose verbs share with v the same synset. Note that frames containing verbs semantically similar but not identical to v do contribute to the probability, thus contributing with an innovation with respect to Alishahi & Stevenson's system. This is particularly important when dealing with few occurrences of v that would not allow further generalization over unseen cases; as noted previously, this is a frequent case in our dataset.

[8] For reasons of space we only displayed the set of semantic properties of one sense for each word.

4 Conclusion

We propose a new system for automatically acquiring SP which integrates frequencies from a Latin treebank with a translation-analogy-enriched version of LWN. Since this research employs a treebank for a less resourced language, it gave us the opportunity to discuss issues related to the size of treebanks for these languages and the integration with other lexical resources, such as wordnets. We showed how methods developed in computational semantics for extant languages and large corpora can be adapted to the special case of a dead language in order to improve the state of the art of its resources. In particular, our approach deals with low frequency items in a novel way by means of a clustering technique which expands the set of seen occurrences that participate in the generalization step, while calculating selectional preferences. In the near future we plan to evaluate this system both against traditional resources such as dictionaries and thesauri, and against corpus data from other sources. Finally, thanks to the shared annotation, running our system on the LDT and the PROIEL treebank would lead to diachronic investigations on Latin syntax and semantics, while at the same time being a computational challenge.

References

[1] A. Alishahi and S. Stevenson. A cognitive model for the representation and acquisition of verb selectional preferences. In *Proceedings of the ACL Workshop on Cognitive Aspects of Computational Language Acquisition. Prague*, pages 41–48, 2007.

[2] D. Bamman and G. Crane. Building a dynamic lexicon from a digital library. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008). Pittsburgh*, 2008.

[3] L. Bentivogli, P. Forner, and and Pianta E. Magnini, B. Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *Proceedings of COLING Workshop on Multilingual Linguistic Resources*, pages 101–108, Geneva, 2004.

[4] C. Brockmann and M. Lapata. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the tenth conference on European chapter of the ACL*, volume 1, Budapest, 2003.

[5] B. McGillivray and M. Passarotti. The development of the *Index Thomisticus* Treebank Valency Lexicon. In *Proceedings of the Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELT&R 2009). Athens*, pages 33–40, 2009.

[6] S. Minozzi. The Latin Wordnet project. In P. Anreiter and M. Kienpointner, editors, *Proceedings of the 15th International Colloquium on Latin Linguistics (ICLL)*, 2009.