Grammatical Role Labeling with Integer Linear Programming

Manfred Klenner
Institute of Computational Linguistics
University of Zurich
klenner@cl.unizh.ch

Abstract

In this paper, we present a formalization of grammatical role labeling within the framework of Integer Linear Programming (ILP). We focus on the integration of subcategorization information into the decision making process. We present a first empirical evaluation that achieves competitive precision and recall rates.

1 Introduction

An often stressed point is that the most widely used classifiers such as Naive Bayes, HMM, and Memory-based Learners are restricted to local decisions only. With grammatical role labeling, for example, there is no way to explicitly express global constraints that, say, the verb "to give" must have 3 arguments of a particular grammatical role.

Among the approaches to overcome this restriction, i.e. that allow for global, theory based constraints, Integer Linear Programming (ILP) has been applied to NLP (Punyakanok et al., 2004) .

We apply ILP to the problem of grammatical relation labeling, i.e. given two chunks.[1] (e.g. a verb and a np), what is the grammatical relation between them (if there is any). We have trained a maximum entropy classifier on vectors with morphological, syntactic and positional information. Its output is utilized as weights to the ILP component which generates equations to solve the following problem: Given subcategorization frames (expressed in functional roles, e.g. subject), and given a sentence with verbs, $\mathcal{V}$ (auxiliary, modal, finite, non-finite, ..), and chunks, $\mathcal{C}$ (np, pp), label all pairs $(\mathcal{V} \cup \mathcal{C}) \times (\mathcal{V} \cup \mathcal{C})$ with a grammatical role[2].

In this paper, we are pursuing two empirical scenarios. The first is to collapse all subcategorization frames of a verb into a single one, comprising all subcategorized roles of the verb but not necessarily forming a valid subcategorization frame of that verb at all. For example, the verb "to believe" subcategorizes for a subject and a prepositional complement ("He believes in magic") or for a subject and a clausal complement ("She believes that he is dreaming"), but there is no frame that combines a subject, a prepositional object and a clausal object. Nevertheless, the set of valid grammatical roles of a verb can serve as a filter operating upon the output of a statistical classifier. The typical errors being made by classifiers with only local decisions are: a constituent is assigned to a grammatical role more than once and a grammatical role (e.g. of a verb) is instantiated more than once. The worst example in our tests was a verb that receives from the maxent classifier two subjects and three clausal objects. Here, such a role filter will help to improve the results.

The second setting is to provide ILP with the correct subcategorization frame of the verb. The results of such an oracle setting define the upper bound of the performance our ILP approach can achieve. Future work will be to let ILP find the optimal subcategorization frame given all frames of a verb.

2 The ILP Specification

Integer Linear Programming (ILP) is the name of a class of constraint satisfaction algorithms which are restricted to a numerical representation of the problem to be solved. The objective is to optimize (minimize or maximize) the numerical solution of linear equations (see the *objective function* in Fig. 1). The general form of an ILP specification is given in Fig. 1 (here: maximization). The goal is to maximize a n-ary function f, which is defined as the sum of the variables $y_i X_i$.

Assignment decisions (e.g. grammatical role labeling) can be modeled in the following way: X_n

[1] Currently, we use perfect chunks, that is, chunks stemming from automatically flattening a treebank.

[2] Most of these pairs do not stand in a proper grammatical relation, they get a null class assignment.

Objective Function:

$$\max f(X_1, \ldots, X_n) := y_1 X_1 + \ldots + y_n X_n$$

Constraints:

$$a_{i1} X_1 + a_{i2} X_2 + \ldots + a_{in} X_n \left(\begin{array}{c} \leq \\ = \\ \geq \end{array} \right) b_i,$$

$i = 1, \ldots, m$
X_i are variables, y_i, b_i and a_{ij} are constants.

Figure 1: ILP Specification

are binary class variables that indicate the (non-) assignment of a constituent c_i to the grammatical function G_j (e.g. subject) of a verb v_k. To represent this, three indices are needed. Thus, X is a complex variable name, e.g. G_{ijk}. For the sake of readability, we add some mnemotechnical sugar and use $G_i v_j c_k$ instead or $\mathcal{S} v_j c_k$ for a constituent c_k being (or not) the subject $\mathcal{S}$ of verb v_j ($\mathcal{S}$ thus is an instantiation of G_i) . If the value of such a class variable $G_i v_j c_k$ is set to 1 in the course of the maximization task, the attachment was successful, otherwise ($G_i v_j c_k = 0$) it failed. y_i from Fig. 1 are weights that represent the impact of an assignment (or a constraint); they provide an empirically based numerical justification of the assignment (we don"t need the a_{ij}). For example, we represent the impact of $G_i v_j c_k$=1 by $\omega_{G_i v_j c_k}$. These weights are derived from a maximum entropy model trained on a treebank (see section 5). b is used to set up numerical constraints. For example that a constituent can only be the filler of one grammatical role. The decision, which of the class variables are to be "on" or "off" is based on the weights and the constraints an overall solution must obey to. ILP seeks to optimize the solution.

3 Formalization

We restrict our formalization to the following set of grammatical functions: subject ($\mathcal{S}$), direct (i.e. accusative) object ($\mathcal{D}$), indirect (i.e. dative) object ($\mathcal{I}$), clausal complement ($\mathcal{C}$), prepositional complement ($\mathcal{P}$), attributive (np or pp) attachment ($\mathcal{T}$) and adjunct ($\mathcal{J}$). The set of grammatical relations of a verb (verb complements) is denoted with G, it comprises $\mathcal{S}, \mathcal{D}, \mathcal{I}, \mathcal{C}$ and $\mathcal{P}$.

The objective function is:

$$max : \mathcal{J} + \mathcal{T} + \mathcal{U} + \mathcal{V} \qquad (1)$$

$\mathcal{J}$ represents the weighted sum of all adjunct attachments. $\mathcal{T}$ is the weighted sum of all attributive PP ("the book in her hand ..") and genitive NP attachments ("die Frau des$_{gen}$ Professors$_{gen}$" [the wife of the professor]). $\mathcal{U}$ represents the weighted sum of all unassigned objects.[3] $\mathcal{V}$ is the weighted sum of the case frame instantiations of all verbs in the sentence. It is defined as follows:

$$\mathcal{V} = \sum_{i}^{|verbs|} \sum_{G \in R_{v_i}} \sum_{j}^{|consts^\star|} w_{G v_i c_j} * G v_i c_j \qquad (2)$$

This sums up over all verbs. For each verb, each grammatical role (R_{v_i} is the set of such roles) is instantiated from the stock of all constituents ($consts^\star$, which includes all np and pp constituents but also the verbs as potential heads of clausal objects). $G v_i c_j$ is a variable that indicates the assignment of a constituent c_j to the grammatical function G of verb v_i. $w_{G v_i c_j}$ is the weight of such an assignment. The (binary) value of each $G v_i c_j$ is to be determined in the course of the constraint satisfaction process, the weight is taken from the maximum entropy model.

$\mathcal{T}$ is the function for weighted attributive attachments:

$$\mathcal{T} = \sum_{i}^{|consts|} \sum_{j\ (i \neq j)}^{|consts|} \omega_{\mathcal{T} c_i c_j} * \mathcal{T} c_i c_j \qquad (3)$$

where $\omega_{\mathcal{T} c_i c_j}$ is the weight of an assignment of constituent c_j to constituent c_i and $\mathcal{T} c_i c_j$ is a binary variable indicating the classification decision whether c_j actually modifies c_i. In contrast to $consts^\star$, $consts$ does not include verbs.

The function for weighted adjunct attachments, $\mathcal{J}$, is:

$$\mathcal{J} = \sum_{j}^{|consts^-|} \sum_{i}^{|verbs|} \omega_{\mathcal{J} v_i c_j} * \mathcal{J} v_i c_j \qquad (4)$$

where $consts^-$ is the set of PP constituents of the sentence. $\omega_{\mathcal{J} v_i c_j}$ is the weight given to a classification of a PP as an adjunct of a clause with v_i as verbal head.

The function for the weighted assignment to the null class, $\mathcal{U}$, is:

$$\mathcal{U} = \sum_{i}^{|consts^\star|} w_{c_i} * \mathcal{U} c_i \qquad (5)$$

This represents the impact of assigning a constituent neither to a verb (as a complement) nor

[3] Not every set of chunks can form a valid dependency tree - $\mathcal{U}$ introduces robustness.

to another constituent (as an attributive modifier). $\mathcal{U}c_i = 1$ means that the constituent c_i has got no head (e.g. a finite verb as part of a sentential coordination), although it might be the head of other c_j.

The equations from 1 to 5 are devoted to the maximization task, i.e. which constituent is attached to which grammatical function and with which impact. Of course, without any further restrictions, every constituent would get assigned to every grammatical role - because there are no co-occurrence restrictions. Exactly this would lead to a maximal sum. In order to assure a valid distribution, restrictions have to be formulated, e.g. that a grammatical role can have at most one filler object and that a constituent can be at most the filler of one grammatical role.

4 Constraints

A constituent c_j must either be bound as an attribute, an adjunct, a verb complement or by the null class. This is to say that all class variables with c_j sum up to exactly 1; c_j then is consumed.

$$\mathcal{U}c_j + \sum_i \sum_G Gv_i c_j + \sum_i \mathcal{T}c_i c_j + \sum_i \mathcal{J}v_i c_j = 1, \ \forall j \tag{6}$$

Here, j is an index over all constituents and G is one of the grammatical roles of verb v_i ($G \in R_{v_i}$).

No two constituents can be attached to each other symmetrically (being head and modifier of each other at the same time), i.e. $\mathcal{T}$ (among others) is defined to be asymmetric.

$$\mathcal{T}c_i c_j + \mathcal{T}c_j c_i \leq 1, \ \forall j, i \tag{7}$$

Finally, we must restrict the number of filler objects a grammatical role can have. Here, we have to distinguish among our two settings. In setting one (all case roles of all frames of a verb are collapsed into a single set of case roles), we can't require all grammatical roles to be instantiated (since we have an artificial case frame, not necessarily a proper one). This is expressed as ≤ 1 in equation 8.

$$\sum_j^{consts^\star} Gv_i c_j \leq 1, \ \forall i, G \in R_{v_i} \tag{8}$$

In setting two (the actual case frame is given), we require that every grammatical role G of the verb v_i ($G \in R_{v_i}$) must be instantiated exactly once:

$$\sum_j^{consts^\star} Gv_i c_j = 1, \ \forall i, G \in R_{v_i} \tag{9}$$

5 The Weighting Scheme

A maximum entropy model was used to fix a probability model that serves as the basis for the ILP weights. The model was trained on the Tiger treebank (Brants et al., 2002) with feature vectors stemming from the following set of features: the part of speech tags of the two candidate chunks, the distance between them in phrases, the number of verbs between them, the number of punctuation marks between them, the person, case and number of the candidates, their heads, the direction of the attachment (left or right) and a passive/active voice flag.

The output of the maxent model is for each pair of chunks (represented by their feature vectors) a probability vector. Each entry in this probability vector represents the probability (used as a weight) that the two chunks are in a particular grammatical relation (including the "non-grammatical relation", NGR) . For example, the weight for an adjunct assignment, $\omega_{\mathcal{J}v_1c_3}$, of two chunks $v1$ (a verb) and $c3$ (a np or a pp) is given by the corresponding entry in the probability vector of the maximum entropy model. The vector also provides values for a subject assignment of these two chunks etc.

6 Empirical Results

The overall precision of the maximum entropy classifier is 87.46%. Since candidate pairs are generated almost without restrictions, most pairs do not realize a proper grammatical relation. In the training set these examples are labeled with the non-grammatical relation label NGR (which is the basis of ILPs null class $\mathcal{U}$). Since maximum entropy modeling seeks to sharpen the classifier with respect to the most prominent class, NGR gets a strong bias. So things are getting worse, if we focus on the proper grammatical relations. The precision then is low, namely 62.73%, the recall is 85.76%, the f-measure is 72.46 %. ILP improves the precision by almost 20% (in the "all frames in one setting" the precision is 81.31%).

We trained on 40,000 sentences, which gives about 700,000 vectors (90% training, 10% test, including negative and positive pairings). Our first experiment was devoted to fix an upper bound for the ILP approach: we selected from the set of subcategorization frames of a verb the correct one (according to the gold standard). The set of licenced grammatical relations then is reduced to the cor-

rect subcategorized GR and the non-governable GR $\mathcal{J}$ (adjunct) and $\mathcal{T}$ (attribute). The results are given in Fig. 2 under F_{corr} (cf. section 3 for GR shortcuts, e.g. $\mathcal{S}$ for subject).

	F_{corr}			F_{coll}		
	Prec	Rec	F-Mea	Prec	Rec	F-Mea
$\mathcal{S}$	91.4	86.1	88.7	89.8	85.7	87.7
$\mathcal{D}$	90.4	83.3	86.7	78.6	79.7	79.1
$\mathcal{I}$	88.5	76.9	82.3	73.5	62.1	67.3
$\mathcal{P}$	79.3	73.7	76.4	75.6	43.6	55.9
$\mathcal{C}$	98.6	94.1	96.3	82.9	96.6	89.3
$\mathcal{J}$	76.7	75.6	76.1	74.2	78.9	76.5
$\mathcal{T}$	75.7	76.9	76.3	73.6	79.9	76.7

Figure 2: Correct Frame and Collapsed Frames

The results of the governable GR ($\mathcal{S}$ down to $\mathcal{C}$) are quite good, only the results for prepositional complements ($\mathcal{P}$) are low (the f-measure is 76.4%). From the 36509 grammatical relations, 37173 were found and 31680 were correct. Overall precision is 85.23%, recall is 86.77% and the f-measure is 85.99%. The most dominant error being made here is the coherent but wrong assignment of constituents to grammatical roles (e.g. the subject is taken to be object). This is not a problem with ILP or the subcategorization frames, but one of the statistical model (and the feature vectors). It does not discriminate well among alternatives. Any improvement of the statistical model will push the precision of ILP.

The results of the second setting, i.e. to collapse all grammatical roles of the verb frames to a single role set (cf. Fig. 2, F_{coll}), are astonishingly good. The f-measures comes close to the results of (Buchholz, 1999). Overall precision is 79.99%, recall 82.67% and f-measure is 81.31%. As expected, the values of the governable GR decrease (e.g. recall for prepositional objects by 30.1%).

The third setting will be to let ILP choose among all subcategorization frames of a verb (there are up to 20 frames per verb). First experiments have shown that the results are between the F_{corr} and F_{coll} results. The question then is, how close can we come to the F_{corr} upper bound.

7 Related Work

ILP has been applied to various NLP problems, including semantic role labeling (Punyakanok et al., 2004), extraction of predicates from parse trees (Klenner, 2005) and discourse ordering in generation (Althaus et al., 2004). (Roth and Yih, 2005) discuss how to utilize ILP with Conditional Random Fields.

Grammatical relation labeling has been coped with in a couple of articles, e.g. (Buchholz, 1999). There, a cascaded model (of classifiers) has been proposed (using various tools around TIMBL). The f-measure (perfect test data) was 83.5%. However, the set of grammatical relations differs from the one we use, which makes it difficult to compare the results.

8 Conclusion and Future Work

In this paper, we argue for the integration of top down (theory based) information into NLP. One kind of information that is well known but have been used only in a data driven manner within statistical approaches (e.g. the Collins parser) is subcategorization information (or case frames). If subcategorization information turns out to be useful at all, it might become so only under the strict control of a global constraint mechanism. We are currently testing an ILP formalization where all subcategorization frames of a verb are competing with each other. The benefits will be to have the instantiation not only of licensed grammatical roles of a verb, but of a consistent and coherent instantiation of a single case frame.

Acknowledgment. I would like to thank Markus Dreyer for fruitful ("long distance") discussions and a number of (steadily improved) maximum entropy models. Also, the detailed comments of the reviewers have been very helpful.

References

Ernst Althaus, Nikiforos Karamanis, and Alexander Koller. 2004. Computing Locally Coherent Discourses. *Proceedings of the ACL. 2004*.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith. 2002. The TIGER Treebank. *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

Sabine Buchholz, Jorn Veenstra and Walter Daelemans. 1999. Cascaded Grammatical Relation Assignment. *EMNLP-VLC'99, the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*.

Manfred Klenner. 2005. Extracting Predicate Structures from Parse Trees. *Proceedings of the RANLP 2005*.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dave Zimak. 2004. Role Labeling via Integer Linear Programming Inference. *Proceedings of the 20th COLING*.

Dan Roth and Wen-tau Yih. 2005. ILP Inference for Conditional Random Fields. *Proceedings of the ICML, 2005*.