Shallow Dependency Labeling

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

Abstract

We present a formalization of dependency labeling with Integer Linear Programming. We focus on the integration of subcategorization into the decision making process, where the various subcategorization frames of a verb compete with each other. A maximum entropy model provides the weights for ILP optimization.

1 Introduction

Machine learning classifiers are widely used, although they lack one crucial model property: they can't adhere to prescriptive knowledge. Take grammatical role (GR) labeling, which is a kind of (shallow) dependency labeling, as an example: chunk-verb-pairs are classified according to a GR (cf. (Buchholz, 1999)). The trials are independent of each other, thus, local decisions are taken such that e.g. a unique GR of a verb might (erroneously) get multiply instantiated etc. Moreover, if there are alternative subcategorization frames of a verb, they must not be confused by mixing up GR from different frames to a non-existent one. Often, a subsequent filter is used to repair such inconsistent solutions. But usually there are alternative solutions, so the demand for an optimal repair arises.

We apply the optimization method Integer Linear Programming (ILP) to (shallow) dependency labeling in order to generate a globally optimized consistent dependency labeling for a given sentence. A maximum entropy classifier, trained on vectors with morphological, syntactic and positional information automatically derived from the TIGER treebank (German), supplies probability vectors that are used as weights in the optimization process. Thus, the probabilities of the classifier do not any longer provide (as usually) the solution (i.e. by picking out the most probable candidate), but count as probabilistic suggestions to a - globally consistent - solution. More formally, the dependency labeling problem is: given a sentence with (i) verbs, $\mathcal{VB}$, (ii) NP and PP chunks[1], $\mathcal{CH}$, label all pairs $(\mathcal{VB} \cup \mathcal{CH}) \times (\mathcal{VB} \cup \mathcal{CH})$ with a dependency relation (including a class for the null assignment) such that all chunks get attached and for each verb exactly one subcategorization frame is instantiated.

2 Integer Linear Programming

Integer Linear Programming 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 (e.g. maximize) a linear equation called the objective function (a) in Fig. 1) given a set of constraints (b) in Fig. 1):

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

$$b) \; a_{i1} X_1 + a_{i2} X_2 + \ldots + a_{in} X_n \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b_i,$$

Figure 1: ILP Specification

where, $i = 1, \ldots, m$ and $X_1 \ldots X_n$ are variables, $y_1 \ldots y_n$, b_i and $a_{i1} \ldots a_{in}$ are constants.

For dependency labeling we have: X_n are binary class variables that indicate the (non-) assignment of a chunk c to a dependency relation G of a subcat frame f of a verb v. Thus, three indices are needed: G_{fvc}. If such an indicator variable G_{fvc} is set to 1 in the course of the maximization task, then the dependency label G between these chunks is said to hold, otherwise ($G_{fvc} = 0$) it doesn't hold. $y_1 \ldots y_n$ from Fig.1 are interpreted as weights that represent the impact of an assignment.

3 Dependency Labeling with ILP

Given the chunks $\mathcal{CH}^+$ (NP, PP and verbs) of a sentence, each pair $\mathcal{CH}^+ \times \mathcal{CH}^+$ is formed. It can

[1] Note that we use base chunks instead of heads.

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 201–204,
Prague, June 2007.

$$\mathcal{M} = \sum_{i}^{|CH|} \sum_{j\ (i \neq j)}^{|CH|} \omega_{\mathcal{T}_{ij}} * \mathcal{T}_{ij} \quad (1)$$

$$\mathcal{A} = \sum_{i}^{|VB|} \sum_{j}^{|PP|} \omega_{\mathcal{J}_{ij}} * \mathcal{J}_{ij} \quad (2)$$

$$\mathcal{V} = \sum_{c}^{|CH^{+}|} \sum_{v}^{|VB|} \sum_{\langle G,f \rangle \in R_v} w_{G_{fvc}} * G_{fvc} \quad (3)$$

$$\mathcal{U} = \sum_{i}^{|VB|} \sum_{j (i \neq j)}^{|VB|} \omega_{U_{ij}} * U_{ij} \quad (4)$$

$$\text{max} : \mathcal{M} + \mathcal{A} + \mathcal{V} + \mathcal{U} \quad (5)$$

Figure 2: Objective Function

stand in one of eight dependency relations, including a pseudo relation representing the null class. We consider the most important dependency labels: subject ($\mathcal{S}$), direct object ($\mathcal{D}$), indirect object ($\mathcal{I}$), clausal complement ($\mathcal{C}$), prepositional complement ($\mathcal{P}$), attributive (NP or PP) attachment ($\mathcal{T}$) and adjunct ($\mathcal{J}$). Although coarse-grained, this set allows us to capture all functional dependencies and to construct a dependency tree for every sentence in the corpus[2]. Technically, indicator variables are used to represent attachment decisions. Together with a weight, they form the addend of the objective function. In the case of attributive modifiers or adjuncts (the non-governable labels), the indicator variables correspond to triples. There are two labels of this type: $\mathcal{T}_{ij}$ represents that chunk j modifies chunk i and $\mathcal{J}_{ij}$ represents that chunk j is in an adjunct relation to chunk i. $\mathcal{M}$ and $\mathcal{A}$ are defined as the weighted sum of such pairs (cf. Eq. 1 and Eq 2. from Fig. 2), the weights (e.g. $\omega_{\mathcal{T}_{ij}}$) stem from the statistical model.

For subcategorized labels, we have quadruples, consisting of a label name G, a frame index f, a verb v and a chunk c (also verb chunks are allowed as a c): G_{fvc}. We define $\mathcal{V}$ to be the weighted sum of all label instantiations of all verbs (and their subcat frames), see Eq. 3 in Fig. 2. The subscript R_v is a list of pairs, where each pair consists of a label and a subcat frame index. This way, R_v represents all subcat frames of a verb v. For example, R of “to believe” could be: $\{\langle \mathcal{S}, 1\rangle, \langle \mathcal{D}, 1\rangle, \langle \mathcal{S}, 2\rangle, \langle \mathcal{C}, 2\rangle, \langle \mathcal{S}, 3\rangle, \langle \mathcal{I}, 3\rangle\}$. There are three frames, the first one requires a $\mathcal{S}$ and a $\mathcal{D}$.

Consider the sentence “He believes these stories”. We have VB=$\{$believes$\}$ and CH^{+} = $\{$He, believes, stories$\}$. Assume R_1 to be the R of “to believe” as defined above. Then, e.g. $S_{213} = 1$ represents the assignment of “stories” as the filler of the subject relation S of the second subcat frame of “believes”.

To get a dependency tree, every chunk must find a head (chunk), except the root verb. We define a root verb j as a verb that stands in the relation $\mathcal{U}_{ij}$ to all other verbs i. $\mathcal{U}$ (cf. Eq.4 from Fig.2) is the weighted sum of all null assignment decisions. It is part of the maximization task and thus has an impact (a weight). The objective function is defined as the sum of equations 1 to 4 (Eq.5 from Fig.2).

So far, our formalization was devoted to the maximization task, i.e. which chunks are in a dependency relation, what is the label and what is the impact. Without any further (co-occurrence) restrictions, every pair of chunks would get related with every label. In order to assure a valid linguistic model, constraints have to be formulated.

4 Basic Global Constraints

Every chunk j from CH ($\neq CH^{+}$) must find a head, that is, be bound either as an attribute, adjunct or a verb complement. This requires all indicator variables with j as the dependent (second index) to sum up to exactly 1.

$$\sum_{c}^{|CH|} \mathcal{T}_{cj} + \sum_{i}^{|VB|} \mathcal{J}_{ij} + \sum_{v}^{|VB|} \sum_{\langle G,f \rangle \in R_v} G_{fvj} = 1, \quad (6)$$

$$\forall j : 0 < j \leq |CH|$$

A verb is attached to any other verb either as a clausal object $\mathcal{C}$ (of some verb frame f) or as $\mathcal{U}$ (null class) indicating that there is no dependency relation between them.

$$\mathcal{U}_{ij} + \sum_{\langle \mathcal{C},f \rangle \in R_i} \mathcal{C}_{fij} = 1, \ \forall i,j (i \neq j) : 0 < i,j \leq |VB| (7)$$

[2] Note that we are not interested in dependencies beyond the (base) chunk level

This does not exclude that a verb gets attached to several verbs as a $\mathcal{C}$. We capture this by constraint 8:

$$\sum_{i}^{|VB|} \sum_{\langle \mathcal{C}, f \rangle \in R_i} \mathcal{C}_{fij} \leq 1, \ \forall j: \ 0 < j \leq |VB| \qquad (8)$$

Another (complementary) constraint is that a dependency label G of a verb must have at most one filler. We first introduce a indicator variable G_{fv}:

$$G_{fv} = \sum_{c}^{|CH^{+}|} G_{fvc} \qquad (9)$$

In order to serve as an indicator of whether a label G (of a frame f of a verb v) is active or inactive, we restrict G_{fv} to be at most 1:

$$G_{fv} \leq 1, \forall v, f, G : 0 < v \leq |VB| \wedge \langle G, f \rangle \in R_v (10)$$

To illustrate this by the example previously given: the subject of the second verb frame of "to believe" is defined as $S_{21} = \mathcal{S}_{211} + \mathcal{S}_{213}$ (with $S_{21} \leq 1$). Either $\mathcal{S}_{211} = 1$ or $\mathcal{S}_{213} = 1$ or both are zero, but if one of them is set to one, then S_{21} = 1. Moreover, as we show in the next section, the selection of the label indicator variable of a frame enforces the frame to be selected as well[3].

5 Subcategorization as a Global Constraint

The problem with the selection among multiple subcat frames is to guarantee a valid distribution of chunks to verb frames. We don't want to have chunk c_1 be labeled according to verb frame f_1 and chunk c_2 according to verb frame f_2. Any valid attachment must be coherent (address one verb frame) and complete (select all of its labels).

We introduce an indicator variable F_{fv} with frame and verb indices. Since exactly one frame of a verb has to be active at the end, we restrict:

$$\sum_{f=1}^{NF_v} \mathcal{F}_{fv} = 1, \ \forall v : \ 0 < v \leq |VB| \qquad (11)$$

(NF_v is the number of subcat frames of verb v)

However, we would like to couple a verb's (v) frame (f) to the frame's label set and restrict it to be active (i.e. set to one) only if all of its labels are active. To achieve this, we require equivalence, namely that selecting any label of a frame is equivalent to selecting the frame. As defined in equation 10, a label is active, if the label indicator variable (G_{fv}) is set to one. Equivalence is represented by identity, we thus get (cf. constraint 12):

$$\mathcal{F}_{fv} = G_{fv}, \ \forall v, f, G : 0 < v \leq |VB| \wedge \langle G, f \rangle \in R_v (12)$$

If any G_{fv} is set to one (zero), then F_{fv} is set to one (zero) and all other G_{fv} of the same subcat frame are forced to be one (completeness). Constraint 11 ensures that exactly one subcat frame F_{fv} can be active (coherence).

6 Maximum Entropy and ILP Weights

A maximum entropy approach was used to induce 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 chunks, the number of intervening verbs, the number of intervening punctuation marks, person, case and number features, the chunks, the direction of the dependency relation (left or right) and a passive/active voice flag.

The output of the maxent model is for each pair of chunks a probability vector, where each entry represents the probability that the two chunks are related by a particular label ($\mathcal{S}, \mathcal{D} \ldots$ including $\mathcal{U}$).

7 Empirical Results

A 80% training set (32,000 sentences) resulted in about 700,000 vectors, each vector representing either a proper dependency labeling of two chunks, or a null class pairing. The accuracy of the maximum entropy classifier was 87.46%. Since candidate pairs are generated with only a few restrictions, most pairings are null class labelings. They form the majority class and thus get a strong bias. If we evaluate the dependency labels, therefore, the results drop appreciably. The maxent precision then is 62.73% (recall is 85.76%, f-measure is 72.46 %).

Our first experiment was devoted to find out how good our ILP approach was given that the correct subcat frame was pre-selected by an oracle. Only the decision which pairs are labeled with which dependency label was left to ILP (also the selection and assignment of the non subcategorized labels).

[3] There are more constraints, e.g. that no two chunks can be attached to each other symmetrically (being chunk and modifier of each other at the same time). We won't introduce them here.

There are 8000 sentence with 36,509 labels in the test set; ILP retrieved 37,173; 31,680 were correct. Overall precision is 85.23%, recall is 86.77%, the f-measure is 85.99% (F$_{pres}$ in Fig. 3).

	F$_{pres}$			F$_{comp}$		
	Prec	Rec	F-Mea	Prec	Rec	F-Mea
$\mathcal{S}$	91.4	86.1	88.7	90.3	80.9	85.4
$\mathcal{D}$	90.4	83.3	86.7	81.4	73.3	77.2
$\mathcal{I}$	88.5	76.9	82.3	75.8	55.5	64.1
$\mathcal{P}$	79.3	73.7	76.4	77.8	40.9	55.6
$\mathcal{C}$	98.6	94.1	96.3	91.4	86.7	89.1
$\mathcal{J}$	76.7	75.6	76.1	74.5	72.3	73.4
$\mathcal{T}$	75.7	76.9	76.3	74.1	74.2	74.2

Figure 3: Pre-selected versus Competing Frames

The results of the governable labels ($\mathcal{S}$ down to $\mathcal{C}$) are good, except PP complements ($\mathcal{P}$) with a f-measure of 76.4%. The errors made with F_{pres}: the wrong chunks are deemed to stand in a dependency relation or the wrong label (e.g. $\mathcal{S}$ instead of $\mathcal{D}$) was chosen for an otherwise valid pair. This is not a problem of ILP, but one of the statistical model - the weights do not discriminate well. Improvements of the statistical model will push ILP's precision.

Clearly, performance drops if we remove the subcat frame oracle letting all subcat frames of a verb compete with each other (F$_{comp}$, Fig.3). How close can F$_{comp}$ come to the oracle setting F$_{pres}$. The overall precision of the F$_{comp}$ setting is 81.8%, recall is 85.8% and the f-measure is 83.7% (f-measure of F$_{pres}$ was 85.9%). This is not too far away.

We have also evaluated how good our model is at finding the correct subcat frame (as a whole). First some statistics: In the test set are 23 different subcat frames (types) with 16,137 occurrences (token). 15,239 out of these are cases where the underlying verb has more than one subcat frame (only here do we have a selection problem). The precision was 71.5%, i.e. the correct subcat frame was selected in 10,896 out of 15,239 cases.

8 Related Work

ILP has been applied to various NLP problems including semantic role labeling (Punyakanok et al., 2004), which is similar to dependency labeling: both can benefit from verb specific information. Actually, (Punyakanok et al., 2004) take into account to some extent verb specific information. They disallow argument types a verb does not "subcategorize for" by setting an occurrence constraint. However, they do not impose *co*-occurrence restrictions as we do (allowing for competing subcat frames).

None of the approaches to grammatical role labeling tries to scale up to dependency labeling. Moreover, they suffer from the problem of inconsistent classifier output (e.g. (Buchholz, 1999)). A comparison of the empirical results is difficult, since e.g. the number and type of grammatical/dependency relations differ (the same is true wrt. German dependency parsers, e.g (Foth et al., 2005)). However, our model seeks to integrate the (probabilistic) output of such systems and - in the best case - boosts the results, or at least turn it into a consistent solution.

9 Conclusion and Future Work

We have introduced a model for shallow dependency labeling where data-driven and theory-driven aspects are combined in a principled way. A classifier provides empirically justified weights, linguistic theory contributes well-motivated global restrictions, both are combined under the regiment of optimization. The empirical results of our approach are promising. However, we have made idealized assumptions (small inventory of dependency relations and treebank derived chunks) that clearly must be replaced by a realistic setting in our future work.

Acknowledgment. I would like to thank Markus Dreyer for fruitful ("long distance") discussions and the (steadily improved) maximum entropy models.

References

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

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*.

Kilian Foth, Wolfgang Menzel, and Ingo Schröder. Robust parsing with weighted constraints. *Natural Language Engineering, 11(1):1-25* 2005.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dave Zimak. 2004. Semantic Role Labeling via Integer Linear Programming Inference. *COLING '04*.