Multiple-step treebank conversion: from dependency to Penn format

Cristina Bosco
Dipartimento di Informatica, Università di Torino
Corso Svizzera 185
10149 Torino - Italia
bosco@di.unito.it

Abstract

Whilst the degree to which a treebank subscribes to a specific linguistic theory limits the usefulness of the resource, the availability of more formats for the same resource plays a crucial role both in NLP and linguistics. Conversion tools and multi-format treebanks are useful for investigating portability of NLP systems and validity of annotation. Unfortunately, conversion is a quite complex task since it involves grammatical rules and linguistic knowledge to be incorporated into the converter program.

The paper focusses on a methodology for treebank conversion which consists in splitting the process in steps corresponding to the kinds of information that have to be converted, i.e. morphological, structural or relational syntactic. The advantage is the generation of a set of parallel treebanks featuring progressively differentiated formats. An application to the case of an Italian dependency-based treebank in a Penn like format is described.

1 Introduction

The usefulness of a treebank can be potentially limited by the degree to which it subscribes to a specific linguistic theory, and when a new annotation is devised which employs a different linguistic framework than a standard, the problem of how to relate the syntactic schemes to one another arises. The increasing availability of multi-format treebanks (e.g. (Bick, 2006)) and the automatic conversion from some formats to others, e.g. (Collins et al, 1999; Bahgat Shehata and Zanzotto, 2006), are attempts to overcome this problem.

The automatic conversion of a treebank plays an important role in NLP and linguistics. First, it increases the exportability of the treebank, making usable tools developed for other resources. Second, it underlies a full check on correctness and consistency of the treebank annotation. Moreover, it is an explicit comparison among formats and linguistic frameworks. Therefore, a conversion is crucial for overcoming the limits imposed by data in formats that realize different grammatical theories to very important activities such as parsing evaluation and comparative testing of the adequateness of a representation for particular linguistic phenomena, languages and/or tasks. For instance, the availability of parallel annotations, and among them one in Penn format, can be of some aid in investigating the irreproducibility of the state-of-the-art results on treebanks or languages other than Penn and English, as empirically demonstrated by, e.g., (Collins et al, 1999) on Czech, (Dubey and Keller, 2003) on German, (Corazza et al, 2004) on Italian.

The paper, first, presents a methodology for the conversion, then an application of the methodology to the conversion of a dependency-based treebank into a Penn-like format, and finally some remarks on the implementation.

2 On the conversion methodology

The conversion of a treebank, annotated with some format A, into format B consists in a simple filtering

Proceedings of the Linguistic Annotation Workshop, pages 164–167,
Prague, June 2007. ©2007 Association for Computational Linguistics

and string manipulation only when A and B both follow the same linguistic framework. Elsewhere the conversion and development of parallel annotations is a challenging task, which involves grammatical rules and linguistic knowledge to be incorporated into the converter programs (see e.g.(Musillo and Sima'an, 2002) (Bick, 2006)). Nevertheless, parallel annotations which employ different linguistic frameworks may serve as a suitable infrastructure for comparisons among them. In fact, the definition of a conversion process is in itself a comparison between A and B, since it involves explicit assumptions about how A and B relate, and a virtually complete and correct mapping which translates every analysis in A into the corresponding analysis in B (Musillo and Sima'an, 2002).

We propose a methodology that consists in organizing the conversion in steps to be performed in cascade. Each step outputs a new annotation format, which differentiates from the previous one only with respect to a single kind of knowledge, e.g. morphological, structural or functional syntactic. The main advantage is in making available a set of parallel annotations for further use too.

In the next part, we describe the application of this methodology to the conversion of the Turin University Treebank (henceforth TUT), which exploits a dependency-based functionally rich annotation, into a Penn-like format.

3 Converting TUT

TUT is a project for an Italian treebank that features a dependency-based annotation following the dependency grammar major tenets (Hudson, 1984). The annotation is centred on a notion of morpho-syntactic-semantic grammatical relation which aims at represent the syntax-semantics interface by means of the Augmented Relational Structure (Bosco, 2004). TUT currently includes 2.000 sentences (see at http://www.di.unito.it/˜tutreeb/) where 200 different dependency relations are annotated. The figure 1 a) shows an example of TUT tree.

Other Italian resources[1] implement, like TUT, particular representation formats and subscribe to specific linguistic frameworks, thus strongly limiting activities such as the application of state-of-the-art parsers and parsing evaluation for this language. The conversion of TUT in a Penn-like format is a crucial step towards the exportability of the resource, but also a first attempt at overcoming these limits by choosing as a further output a format of widespread use in training, testing and evaluating. Moreover, since the process is fully deterministic, even if currently applied on a small corpus, the conversion is in itself a preliminary validation of the resource and a demonstration that TUT annotation is expressive at least as Penn.

In the next sections, we show the translation of dependency into constituency trees and the management of differences in PoS tagging, structural syntactic, and syntactic-semantic relations faced during the conversion. For detailed information about the conversion of specific linguistic phenomena see at http://www.di.unito.it/˜tutreeb/noteparallele.zip.

3.1 First step: morphology

Since Italian is inflectionally richer than English, TUT PoS tagset is richer than that of Penn (see at http://www.di.unito.it/˜tutreeb/syntcat-22-7-02.doc), but we reduced it including only information that Penn makes explicit too, as usual in similar cases, see e.g. (Collins et al, 1999) and http://www.coli.uni-sb.de/sfb378/negra-corpus/. The major differences with respect to Penn concern tags of Verbs, which include fine-grained temporal information and are organized in three classes (Modal, Auxiliary, Main), rather than two like in Penn (Modal and non-Modal). Moreover, a fine-grained variety of Adjectives and Pronouns enables the recovery of information such as e.g. the owner of an object (possessive Adjective).

The output of this first step includes compact tags where features are expressed by short strings, like in (Collins et al, 1999). The following are examples of TUT PoS native vs reduced tags: for a common Noun 'nome' is (NOME NOUN COMMON M SING) reduced in (NOU˜CS); for the main infinite Verb 'entrare' is (ENTRARE VERB MAIN INFINITE PRES INTRANS) reduced in (VMA˜IN).

3.2 Second step: structural syntax

The main issue in this step is the conversion of dependency trees into Penn-like trees, i.e.

[1] Two other larger Italian treebanks exist: Venice Italian Treebank (VIT) (Delmonte, forthcoming) and Italian Syntactic Semantic Treebank (ISST) (Barsotti et al, 2001)

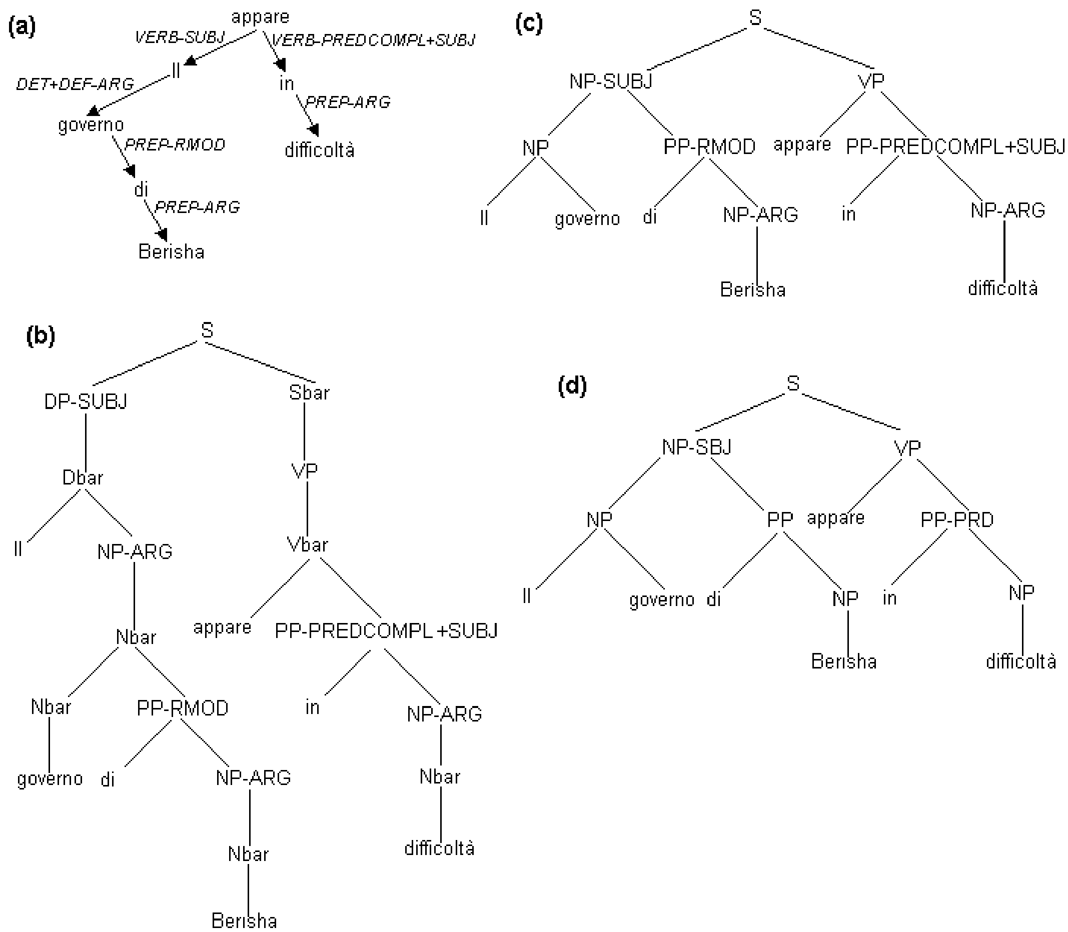

Figure 1: TUT (a), Constituency-TUT (b), Augmented-Penn (c) and Penn-like (d) representations of sentence ALB-4 ”Il governo di Berisha appare in difficoltà” (The government of Berisha appears in trouble)

constituency structures implementing a minimal projection strategy. It is approached in two substeps: by first converting the TUT trees into a standard linguistically motivated Xbar form (i.e. Constituency-TUT), and then into Penn format (i.e. Augmented-Penn), but both including the functional syntactic information as TUT, i.e. the grammatical relations (annotated on constituents).

Constituency-TUT is a TUT-oriented constituency-based annotation that introduces in TUT trees the types of the multiple words syntactic units (e.g. VP and S). Each terminal category X corresponds to a word of a TUT tree, and projects into non-terminal nodes, namely the intermediate (Xbar) and maximal (XP) projections of X. The distinction between complements and adjuncts is here structurally marked.

Augmented-Penn instead features a format structurally isomorphic to Penn, but more functionally annotated. It applies to the Constituency-TUT structures the minimal projection strategy[2], and manages the smoothing of structures conceptually different in TUT and Penn, i.e. those of Determiners, auxiliary Verbs and relative clauses. In figure 1 you can see the same sentence in TUT, Constituency-TUT, Augmented-Penn and Penn format.

The conversion from dependency to constituency is not affected by the typical problem of non-projective structures, since TUT represents them through projective structures exploiting null elements. In dependency TUT, empty nodes also mark dropped subjects, and Constituency-TUT exploits

[2]Each terminal category projects only when the constituent includes more than one word

null elements for marking subjects which occur in non standard position with respect to the Verb (i.e. extraposed).

3.3 Third step: syntactic-semantic relations

While Penn features a description of relations based only on a single component, TUT features an explicit, systematic annotation of three components in each relation. Moreover, Penn includes a lower number of values for each component than TUT[3] and in various cases the Penn tags do not enable fine-grained distinctions as TUT.
We applied the original Penn tags that can be meaningful for Italian looking for correspondences between TUT and Penn relations (e.g. using the relation LOC for all TUT LOC+ relations)[4]. Nevertheless, the multi-step methodology makes available also a Penn-like format almost functionally rich as TUT, i.e. Augmented-Penn[5].

4 The converter

The five modules of the converter are: M_{reduc} for the reduction of PoS tags; M_{ctu} which converts in the Constituency-TUT format; M_{augp}, which converts Constituency-TUT in Augmented-Penn; M_{pen}, which takes Augmented-Penn and outputs Penn; M_{par} that generates the parenthetical notation of the output.
M_{ctu} manages the conversion from dependency to constituency by implementing the algorithm in (Xia, 2001). It recovers the types of phrases that (the grammatical category of) each node of the dependency tree projects by using the linguistic knowledge stored in dedicated tables.
The converter follows a lowest attachment strategy, i.e. the projection of a dependent attaches to a projection of its head as lowly as possible, but, in contrast with (Xia, 2001), it pursues a maximal rather than minimal projection heuristics, i.e. a category always projects into intermediate and maximal projections.

[3]While Penn annotates 2 morpho-syntactic, 11 syntactic and 7 semantic relations, TUT features 40 morpho-syntactic, 55 functional-syntactic and 88 semantic items for building relations.

[4]The conversion from NEGRA to Penn maintains instead the NEGRA relations, see at http://www.coli.uni-sb.de/sfb378/negra-corpus/.

[5]The relations linking terminal nodes encompassed in a single constituent in Augmented-Penn are deleted during the conversion in this latter format.

5 Conclusions

The methodology for treebank conversion here presented splits the process in steps, which correspond to the kinds of annotated linguistic knowledge that have to be converted. Since each step outputs a new annotation format, the advantage is the generation of set of parallel treebanks.
The application of the methodology in the conversion from a small Italian dependency-based treebank to a Penn like format is described.

References

A. Bahgat Shehata and F.M. Zanzotto. 2006. A dependency-based algorithm for grammar conversion. *Proc. of LREC '06*

F. Barsotti and R. Basili et al. 2001. *The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation.* Kluwer, Dordrecht.

E. Bick. 2006. Turning a dependency-based treebank into a PSG-style constituent treebank. *Proc. of LREC 06.*

C. Bosco. 2004. *A grammatical relation system for treebank annotation.* PhD thesis, University of Torino.

M. Collins, J. Hajic, L. Ramshaw, and C. Tillmann. 1999. A statistical parser of Czech. *Proc. of ACL'99.*

A. Corazza, A. Lavelli, G. Satta, and R. Zanoli. 2004 Analyzing an Italian treebank with state-of-the-art statistical parser. In *Proc. of TLT-2004.*

R. Delmonte. forthcoming. *Strutture sintattiche dall'analisi computazionale di corpora di italiano* Franco Angeli, Milano, Italy.

A. Dubey and F. Keller. 2003. Probabilistic parsing for German using sister-head dependencies. *Proc. of ACL'03.*

R. Hudson. 1984. *Word Grammar.* Basil Blackwell, Oxford and New York.

G. Musillo and K. Sima'an. 2002. Towards comparing parsers from different linguistic frameworks. An information theoretic approach. *Proc. of Workshop Beyond PARSEVAL.*

F. Xia. 2001. *Automatic grammar generation from two different perspectives.* PhD thesis, University of Pennsylvania.