Tree distance and other variants of evalb

Martin Emms Dept. of Computer Science, Trinity College, Dublin Martin.Emms@tcd.ie

The `evalb` *measures of parser performance basically treat gold-standard and parser-generated trees* $\mathcal{G}$ *and* $\mathcal{T}$*, as* sets of labelled spans, $\mathcal{G}^S$ *and* $\mathcal{T}^S$*, and quantify the similarity of these sets by precision and recall scores, often combined into* $F1$*, itself equivalent (as shown below) to the* Dice *similarity measure.*

tree-distance *[1] is an alternative to this way of proceeding, which treats trees in their own right, as representations of ancestry and linearity, rather than via the projection into sets of labelled spans. This work applies this measure to parser evaluation, where it has never been used. Some other variants of the standard* `evalb` *procedures are also considered.*

T-mappings and E-mappings

tree-distance and the `evalb` measures can all be subsumed under the perspective of a cost assigned to a mapping. Given any *partial*, one-to-one mapping between two trees $\sigma : \mathcal{G} \mapsto \mathcal{T}$ define

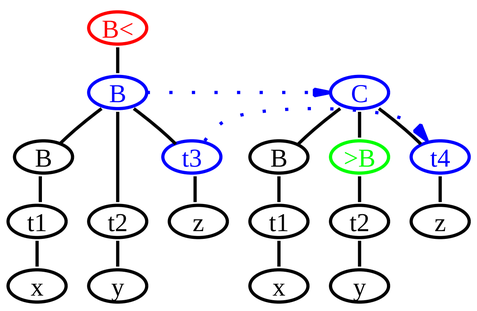
sets for σ $\begin{cases} \text{Deletions } \mathcal{D} = \{n \in \mathcal{G} : n \notin dom(\sigma)\} \\ \text{Insertions } \mathcal{I} = \{n \in \mathcal{T} : n \notin ran(\sigma)\} \\ \text{Swaps } \mathcal{S} = \{n \in \mathcal{G} : label(n) \neq label(\sigma(n))\} \\ \text{Matches } \mathcal{M} = \{n \in \mathcal{G} : label(n) = label(\sigma(n))\} \end{cases}$ *cost* of σ is $D + I + S$

Two differing sets of further conditions on mappings essentially distinguish the tree-distance approach from the standard approach:

T-mappings $\begin{cases} \text{(T1) } \textit{preserve left-to-right order} \\ \text{(T2) } \textit{preserve ancestry} \end{cases}$ E-mappings $\begin{cases} \text{(E1) } \textit{preserve node labels} \\ \text{(E2) } \textit{preserve lexical spans} \end{cases}$

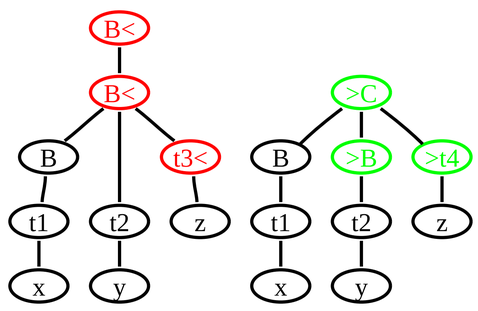
Given two trees, there are many T (and E) mappings between them. The *T-distance* between two trees is the cost of the *minimum-cost* mapping. Similarly for *E-distance*

a least-cost T-mapping:



1 deletion, 1 insertion, 2 swaps
Cost = 4

a least-cost E-mapping



3 deletions, 3 insertions
Cost = 6

Normalisations

The above-defined costs will tend to grow with the size of trees compared, leading to the question of normalisation.

A technicality: the standard `evalb`-quantities refer to what we will call the *roof* of a tree: the nodes which are not terminal or pre-terminal. Let $\hat{\cdot}$ refer to restriction to the *roof* part of trees.

3 standard `evalb` measures are *labelled rec.* $R : \hat{M}/\hat{G}$, *labelled prec.* $P : \hat{M}/\hat{T}$ and their $F1$ combination: $2RP/(R+P)$.

By substituting R and P into F1, and simplifying, F1 can be shown to be equivalent to the *Dice* formulae for comparing 2 sets

$$\begin{aligned} F1 &= (2 \times \hat{M}/\hat{G} \times \hat{M}/\hat{T})/(\hat{M}/\hat{G} + \hat{M}/\hat{T}) \\ &= 2\hat{M} \times (1/\hat{G}\hat{T})/(1/\hat{G} + 1/\hat{T}) \\ &= 2\hat{M} \times (1/\hat{G}\hat{T})/((\hat{G} + \hat{T})/\hat{G}\hat{T}) \\ &= 2\hat{M}/(\hat{G} + \hat{T}) \end{aligned}$$

Thus the standard R, P, and $F1$ are just different *normalisations* of the *match* score of a least-cost E-mapping. There is one further natural alternative, *Jaccard*, which normalises by $\hat{G} \cup \hat{T}$:

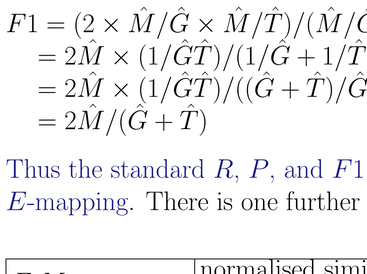
E Measure	normalised similarities $1 \ldots 0$	normalised distances $0 \ldots 1$
labelled rec. R *labelled prec.* P	$\hat{M}/\hat{G}$ $\hat{M}/\hat{T}$	$\hat{D}/\hat{G}$ $\hat{I}/\hat{T}$
$F1$ $= E\ Dice$	$2RP/(R+P)$ $= 2\hat{M}/(\hat{G}+\hat{T})$	 $(\hat{D}+\hat{I})/(\hat{G}+\hat{T})$
E Jaccard	$\hat{M}/(\hat{G} \cup \hat{T})$ $= \hat{M}/(\hat{G}+\hat{T}-\hat{M})$	$(\hat{D}+\hat{I})/(\hat{G} \cup \hat{T})$ $= (\hat{D}+\hat{I})/(\hat{G}+\hat{T}-\hat{M})$

For each normalised similarity s there is normalised distance d, such that $s + d = 1$, using the fact that under E1/E2, $\hat{M} = \hat{G} - \hat{D} = \hat{T} - \hat{I}$, and $|\hat{G} \cup \hat{T}| = \hat{G} + \hat{T} - \hat{M}$
The *whole-tree* variant includes the pre-terminals

$TMeasure$	*normalised distances* $0 \ldots 1$	*derived similarities* $1 \ldots 0$
$TDice$	$\dfrac{D+I+S}{(G-W)+(T-W)}$	$1 - \dfrac{D+I+S}{(G-W)+(T-W)}$
$TJaccard$	$\dfrac{D+I+S}{D+S+M+I-W}$	$1 - \dfrac{D+I+S}{D+S+M+I-W}$

Unlike an E-mapping, a T-mapping is possible between trees with different lexical yields. In a least-cost T-mapping $\sigma : \mathcal{G} \mapsto \mathcal{T}$, lexical items are usually, though not always, mapped to each other. The normalisations diminish the significance of large numbers of word matches.

Differences: a priori



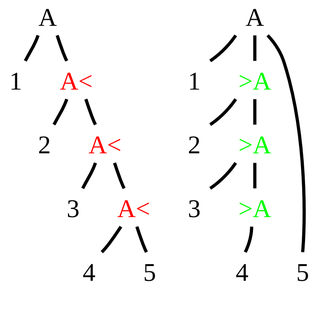
not every E-mapping is a T-mapping, because span-preservation does not imply ancestry preservation.

Unary branching is the hallmark of cases where E-mappings are not T-mappings

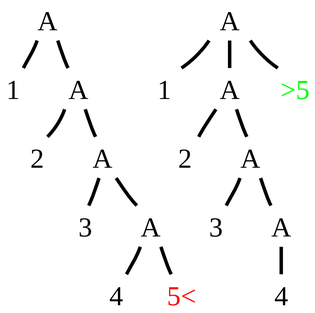
Apart from these, an E-mapping will be a T-mapping, and with greater than or equal cost: thus one would expect E-similarities to be generally *lower than* T-similarities

A case where the span-preserving E-mapping incurs higher costs than the ancestry-preserving T-mapping is the often-noted over-penalisation of attachment errors by the E-measure.

E-mapping, cost 6, *E Dice similarity* 0.25.



T-mapping, cost 2, *T Dice similarity* 0.75



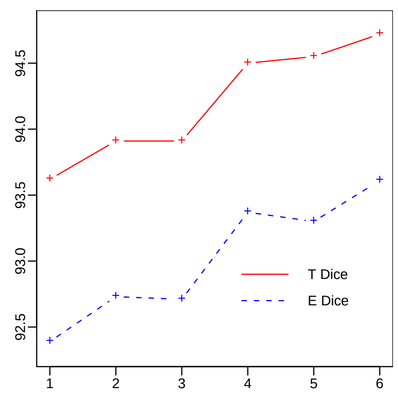
Differences: empirical I

Do parsing systems rank differently on variants of evalb ?

variants: T vs E, Dice vs Jaccard, whole vs roof, macro vs micro averaging

parsers: *Collins 1/2/3*: the 3 models of [2], *Charniak*: the max. ent. parser of [3], *Petrov 5/6*: the 5 and 6 split-merge cycle versions of the parser of [4].

evalb vs Tree-distance

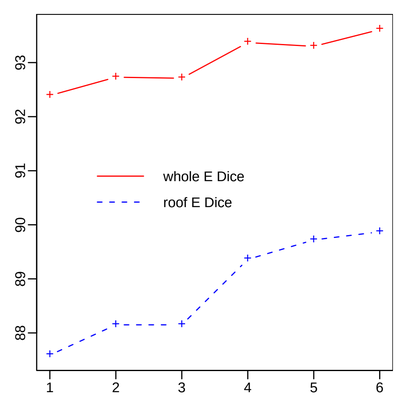


Plot shows E and T scores (whole-tree, macro-averaged). x-axis 1-3 = Collins 1/2/3, 4 = *Petrov 5*, 5 = *Charniak*, 6 = *Petrov 6*. Same in later plots

T-Dice: *Petrov 5* < *Charniak*
E-Dice: *Charniak* < *Petrov 5*

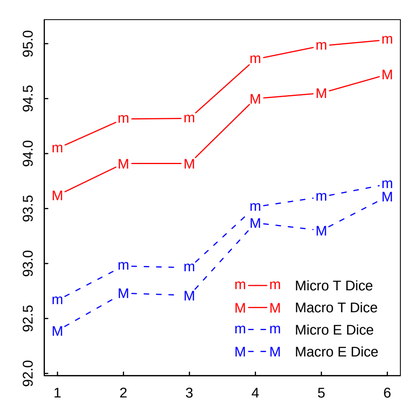
In line with expectation, the T scores are higher than the E scores.

Whole vs Roof Tree



Comparing *E Dice* scores for whole trees and roof trees reverses the *Charniak* < *Petrov 5* ordering. The effect persists with the Jaccard normalisation.

Micro vs Macro averaging



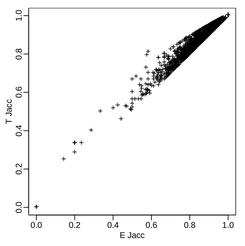
Comparing micro and macro-averaged *E Dice* scores reverses the *Charniak* > *Petrov 5* ordering.

Dice vs Jaccard normalisation: little impact on ranking, though large impact on absolute value: Jaccard is about 7% *lower*. For the micro-averaged *E score*, changing from Dice to Jaccard did change the ranking.

Differences: empirical II

for a fixed system, do the variants of evalb rank the parse*s* differently ?

T vs E
when E-scores are plotted against T-scores, a smeared-out band results, which suggests the E-ranking and T-ranking of parses will differ.
The *kendall-tau* measure [5] of the rank difference between the E and T rankings is 4-5%, for all the parsers, and with either normalisation (indicates how often the rankings permute a pair)

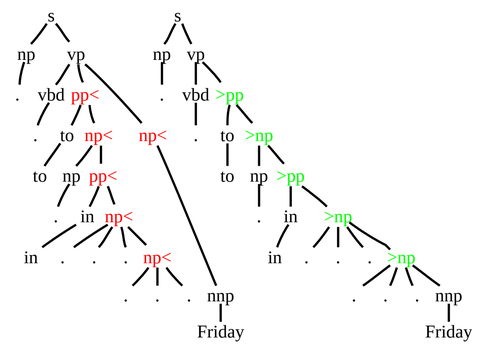


Sentence 159 in the Section 23 test set was

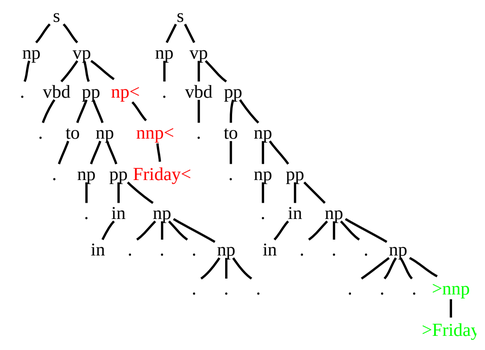
> *Vincent Bajakian manager of the $ 1.8 billion Wellington Fund added to his positions in Bristol-Myers Squibb Woolworth and Dun & Bradstreet Friday*

and in the reference parse *Friday* is attached high, in the *Petrov 5* parse it is attached low. The E-ranking is 504 places lower than the T-ranking.

the E-mapping



the T-mapping



Conclusions

The *ancestry*-preservation of tree-distance is a natural alternative to the *span*-preservation the evalb scores. Tree distance can give more intuitive outcomes than evalb for certain attachment errors, and the ranking of parsing systems under tree distance can differ from that by evalb. Amongst the further issues to explore are: the case where empty nodes are included, the relationship to the *Leaf-Ancestor* metric, and application to other treebanks.

Software

software for calculating tree distance is available at `www.cs.tcd.ie/Martin.Emms/tdist`

References

[1] K.Zhang and D.Shasha, "Simple fast algorithms for the editing distance between trees and related problems," *SIAM Journal of Computing*, vol. 18, pp. 1245–1262, 1989.

[2] Michael Collins, "Head-driven statistical models for natural language parsing," *Computational Linguistics*, vol. 29, no. 4, pp. 589–637, 2003.

[3] Eugene Charniak, "A maximum-entropy-inspired parser," in *Proceedings of NAACL 2000*, 2000, pp. 132–139.

[4] Slav Petrov, Leon Barret, Romain Thibaux, and Dan Klein, "Learning accurate, compact, and interpretable tree annotation," in *Proceedings of COLING/ACL 2006*, 2006, pp. 433–440.

[5] S.Siegel and N.J.Castellan, *Non-Parametric Statistics for the Behavioural Sciences*, McGraw-Hill, 1988.