Tool Performance

The quality of output from tools for automatic analysis/processing of natural language hardly ever reaches 100% and depends on several factors:

- The annotation task (e.g., part-of-speech tagging vs. parsing)
- The input data: For in-domain input performance is reasonably well (comparable to results from literature). Out-of-domain input likely results in a drop of performance.

This might result in trust issues for potential users, especially when the tool is applied as an intermediate step in a processing chain or in a new out-of-domain setting. Furthermore, workload for the user might increase by a need for additional evaluation to find the best tool for the task or extensive correction steps after processing to find and correct errors.

Preserving Tool Confidence

Many automatic tools are internally aware of a relative reliability of their output since they make use of probabilities and forced guessing to decide on a single prediction and unexpected input, from which a reliability estimation can be derived. In most cases this information is discarded after deciding on a single prediction and never included in a tool’s output. In our opinion this represents a serious loss of valuable meta-information.

Confidence as Annotation Layer

We suggest to include a tool’s confidence estimations directly in its output as additional (meta-)annotation layers. This allows for a series of straightforward applications:

- Visualizations can easily use this information to present visual distinctions or filters based on different levels of confidence.
- When using sophisticated corpus-query tools the user is able to restrict queries to annotations with desired reliability via search constraints.

⇒ Ultimately the availability of those values as annotations should aid users in finding interesting data wrt to a certain range or level of confidence.

Interpretation of Confidence Values

Unprocessed usage of raw confidence values in cases where they are already available (e.g., probabilities produced by BitPar [1] for the left trees in above figure) faces certain limitations in terms of usability:

- Range of possible values can be vastly different across and within tools
- Granularity of analysis can differ between tools (e.g. confidence for individual arcs in a dependency parse versus the entire tree)

⇒ Currently real comparability is only possible for confidence values produced by a single tool for the same kind of decisions, such as n-best lists.

We therefore propose to normalize to a simple scale to allow for a more general and comparable interpretation of individual confidence estimations as well as a basic comparability between different tools. As such a tool should project its confidence estimation into classic probability values in the closed interval $[0, 1]$ (as a confidence scale from ‘pure guesswork’ to ‘being sure’).

Advantages of Transparent Confidence

It is important to keep in mind that the availability of confidence estimations (obtained either internally from a tool or externally by comparison of multiple outputs) does not increase the quality of an annotation as such. But it rather boosts usability of large automatically annotated datasets by:

- Raising awareness wrt reliability in the first place
- Helping users to assess if an analysis or part of it is sufficiently reliable, e.g. for a specific downstream task

Thus transparent confidence values can foster the application of state-of-the-art tools on out of domain data, when used in compositional architectures and in related fields such as the Digital Humanities.

www.ims.uni-stuttgart.de