

Toward an Integrated Evaluation Framework

Contribution for the FlaReNet launching event (Vienna, February 12-13, 2009)

Bernardo Magnini
FBK-irst, Trento, Italy

In the last years we assisted to an increasing offer of evaluation campaigns in the area of language technologies. Roughly, we can individuate two types of such campaigns: task-oriented evaluations and application-oriented evaluations.

In *Task-oriented evaluation* a single task is evaluated independently of the final application scenario. This approach tends to maximize task performance and to reuse methodologies (e.g. machine learning) through tasks. Examples of successful task-oriented evaluations are named entities recognition, semantic role labeling and word sense disambiguation. While task-oriented evaluation has generated an impressive number of initiatives, in several cases it is still difficult to understand the impact of a single component in the final scenario.

In *Application-oriented evaluation*, the overall application scenario is evaluated independently of the intermediate components involved in the process. Examples of successful application-oriented evaluations are question answering, summarization and machine translation. In application-oriented evaluation the focus case it is usually difficult to understand the role of single components. As a consequence, the approach tends to maximize global performance.

We suggest an *integrated evaluation framework* where single components are not evaluated per se, but rather for their contribution to a global application. The ideal infrastructure for integrated evaluation would be a network of web services, where each web service serves a specific component. The main expected benefits are that (i) components which contribute most will be rewarded, in term of interests, this way fostering new research directions; (ii) new metrics will be developed to evaluate single components in complex architectures; (iii) a larger amount of ablation tests both for components and for resources will be available, as well as fine grained quantitative and qualitative analysis; (iv) new methodologies for faster prototyping of final applications will be developed.

According with the above considerations, we suggest the following roadmap toward integrated evaluation (5 years):

- Develop shared communication protocols for single-task components;
- Support interoperability of single-task components within global applications;
- Set up a web infrastructure based on web services on the base of shared communication protocols;
- Promote the use of ablation tests in current and future evaluation initiatives, both for resources and for tools.