Resources in Cross-Language Information Retrieval.
Position Paper.

Julio Gonzalo, UNED (Madrid)

Survey on resources

Questionnaire

- Resources used for CLEF runs (provider, availability, coverage)
- Adaptation/enrichment for CLEF (formatting, correction, filtering, enrichment)
- Strenghts and weaknesses
- Key issues for future CLTR resources

Survey (1): Resources

- Monolingual dict. (1)
- Bilingual dict. (8) web or publishing companies
- Multilingual lexical databases (3)
- Bilingual corpora (2 + 2 mined from the web)
- Multilingual corpora (1 Eng-Sp-Fr, 1 Eng-Fr-Sp-It-Ge)
- Machine Translation systems (2)
- Morphological analysers + taggers (5)

Survey (2): Strengths and weaknesses

- Some dictionaries and MT systems freely available on the web ... But usually of low quality and coverage.
- Some dictionaries with good coverage (90% CLEF queries), context/domain information, etc. ... But usually restricted for access or expensive for research purposes.

Survey (3): Key issues for the future

- Availability (9)
  - Of resources, including dictionaries, corpora, indexed collections to evaluate retrieval modules, resources to cope with named entities, etc.
  - Of indexing tools for different languages (e.g. stemmers)
- Quality, coverage, price, maintenance, good markup, semantic relations and

Funding for LE resources in Europe

- Resource building projects funded in past EC programs... but difficult in the Fifth framework.
- Heterogeneous consortia (universities + software companies + publishers) ➔ copyright and distribution problems. Dict. providers vs. researchers
- Little emphasis on evaluation of final products, diverse quality and coverage.

What could be done in CLEF?

- Provide evaluation of resources in a standard retrieval framework
- Provide evaluation of CLTR strategies with a fixed, standard resource.
- Larger set queries for fair evaluation of coverage (perhaps via known item retrieval using titles against documents in a bilingual news collection).
- Support from ELRA.