UTACLIR @ CLEF 2002: Towards unified translation process

CLEF Workshop 2002

Rome, 19. - 20.9.2002

Eija Airio, Heikki Keskustalo,
Turid Hedlund, Ari Pirkola
Department of Information Studies
University of Tampere, Finland

Cross-language information retrieval

- Goal
  - query expressed in one language retrieves documents in multiple languages
- Multilingual, bilingual and even monolingual retrieval tasks can be seen as steps towards this goal
- A unified translation process for multiple language pairs as a step towards wide-ranging multilingual information retrieval : UTACLIR system

Background of UTACLIR

- Bilingual processes for CLEF 2000 and 2001: Swedish - English, German -English and Finnish – English
- The idea of UTACLIR is based on translating topic words one by one, and then combining the translations into the query
- C programs on Solaris 7

UTACLIR 2000 and 2001: basic principles

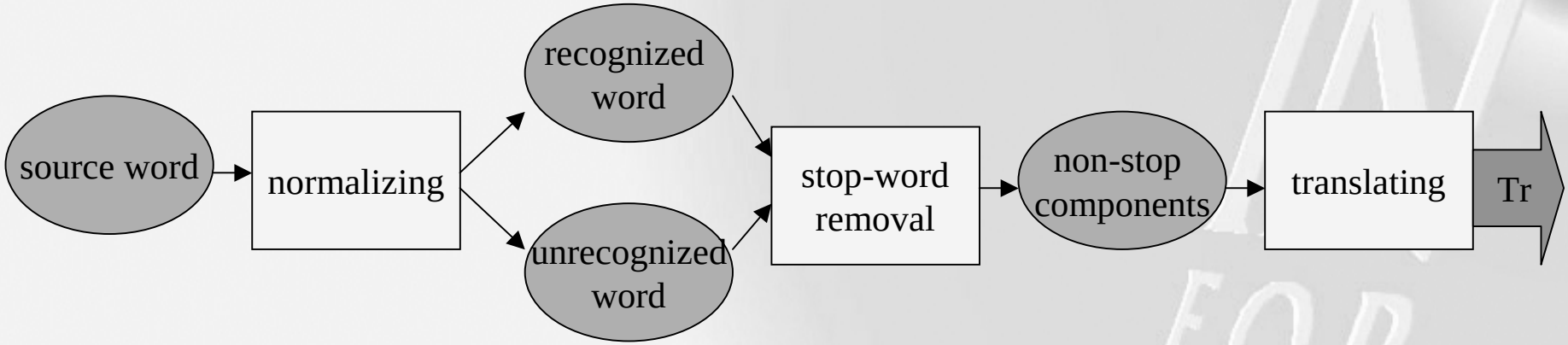
- topic words normalized by a morphological analyzer
- source stop word removal
- translation
- translated words normalized
- target stop word removal
- approximate string matching techniques applied for untranslatable words
- structuring of queries using the synonym operator
- splitting of untranslatable compounds

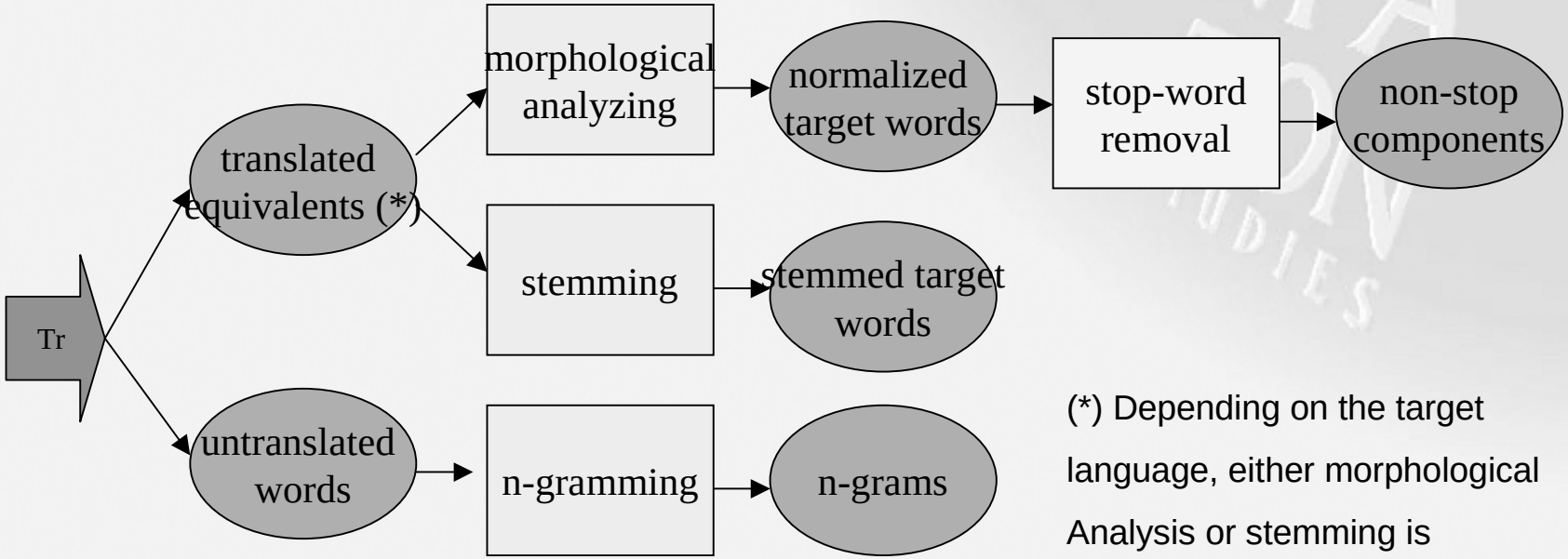
The new UTACLIR process

- operates on Solaris 7, progammed in C
- consists of library archives containing general and resource specific functions
- the same basic principles as in the earlier versions
- the same process for all the languages

The new UTACLIR process(cont.)

- the user gives the codes expressing the source and the target language
- the system uses external linguistic resources depending on the codes
- possible to add parallel resources





(*) Depending on the target language, either morphological Analysis or stemming is performed

CLEF runs 2002 with the new UTACLIR

- *The dictionary and the normalizers used in the runs*

- Motcom GlobalDix multilingual translation dictionary (18 languages, total number of words 665 000)
- Morphological analysers FINTWOL, GERTWOL and ENGTWOL
- Stemmers for Spanish and French, by Zprise
- A stemmer for Italian, by the University of Neuchatel

CLEF runs 2002 (cont.)

- the runs were done with a beta-version
  - splitting of compounds was not yet implemented
  - n-grams methods applied only in English – German run
  - implementing of the Italian and Spanish dictionaries was not ready

- probably better result when the new UTACLIR is ready

Our results in CLEF 2002

	Average precision %
English – Finnish	20.2
English – French	23.9
Multi-lingual I	16.4
Multi-lingual II	11.7

-an additional English – Finnish run to clarify the effect of the dictionary on the result

- the larger MOT dictionary with 110 000 Finnish – English entries was utilized
- the result was 32.6%, 61.4 % better then the original CLEF-result

Result merging vs. index merging ?

- should we research index merging or result merging?
- problems of result merging:
  - result lists are not comparable
  - differs from the Internet approach

Problems in multilingual indexing

- how to build a merged index?
  - the indexing program should call multiple morphological analyzers and stemmers – how?
- shall we use language resources for building the indexes?
- how would we do withouth normalized indexes?
  - the translation produces normalized queries – how can we match them with the unnormalized indexes?

Conclusions

- the existence of unified translation systems, such as UTACLIR, can be seen as a precondition for realistically carrying out CLIR for a large number of languages
- UTACLIR system has proved its competitiveness in translating multiple language pairs

Conclusions (cont.)

- the result merging vs. index merging is still a challenge for cross-language information retrieval
- examples of possible goals:
  - develop translation systems for Internet
    - concentrate on translating and forget result merging (merged indexes are needed for testing the systems)
  - develop systems for environments where the indexes are separate
    - result merging is the goal