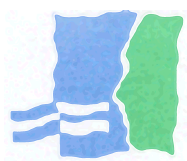
MaTrEx: a Quick Overview

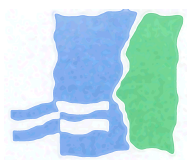
Nicolas Stroppa
nstroppa@computing.dcu.ie

Dublin City University, School of Computing, NCLT

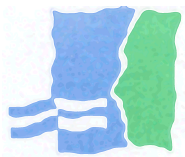
June 15, 2007



Outline



Outline



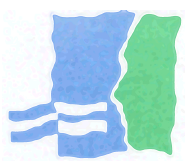
The MaTrEx Project

Goals

- Re-writing prototypes to build re-usable systems
- Synchronizing development, avoiding duplicates
- Scaling to larger datasets (for evaluations, etc.)
- Working as a team: learning from each other, benefiting from others' skills, etc.

Properties in mind

- Modularity, so that anyone can add new features and adapt the system to its own needs
- Maintainability (short turnovers...)
- Efficiency, to deal with large datasets and to allow for more experiments, etc.



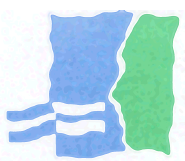
MaTrEx History

System's core: re-implementation of prototypes

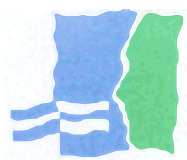
- Marker-Based EBMT (Nano Gough, PhD 2005, CL 2003)
- Hybrid Data-Driven MT (Declan Groves, PhD 2007, MT 2006)

Evaluations

- OpenLab 2006
- NIST 2006
- IWSLT 2006



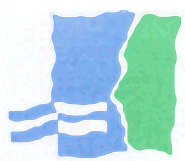
Outline



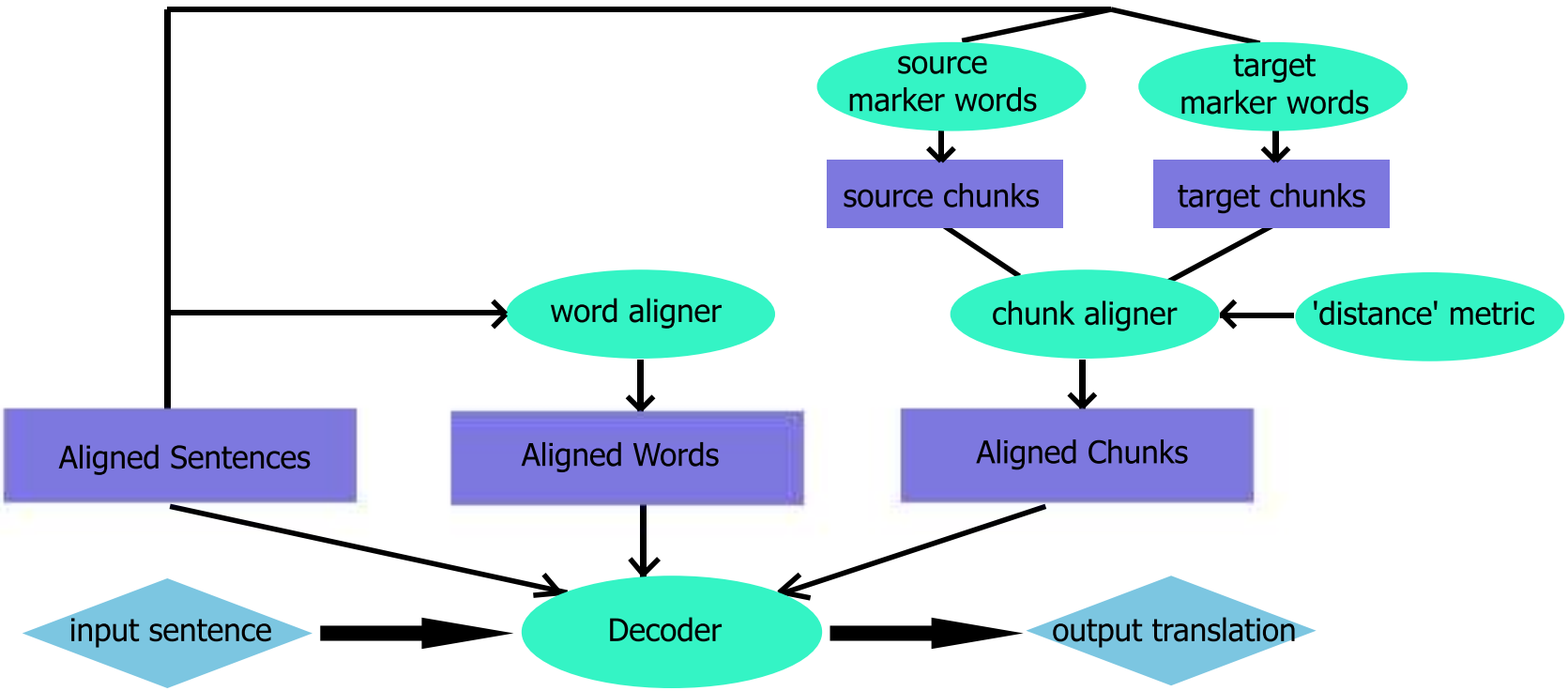
MaTrEx: A Hybrid EBMT/SMT System

Overview of the system

- A word alignment component (GIZA++)
- A chunking component
- A chunk alignment component
- Two phrase alignment components:
  - "SMT"-style phrase aligner (standard phrase extraction from GIZA++ alignments)
  - "EBMT"-style phrase aligner (phrases are extracted from (i) the chunker and (ii) the chunk aligner)
- A minimum-error rate training component (wrapper around MOSES)
- A decoder (wrapper around MOSES)
- A case and punctuation restoration component



MaTrEx: architecture



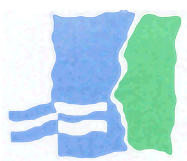
Chunking and Chunk alignment

Several chunking strategies

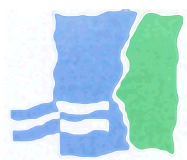
- Marker-based chunking
  - surface chunking based on marker words
- Treebank-based chunking
  - learner trained on annotated data extracted from treebanks

Several chunk alignment strategies

- Edit-distance-like alignment
- Edit-distance-(with jumps)-like alignment
- IBM model-1-like alignment



Outline



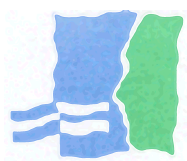
Recent addition (1): Word Packing

Word-packing and bilingually-motivated tokenisation (Ma et al., ACL 2006)

- Word alignment relies on segmenting sentences into basic units ("words")
- Word packing: packing (consecutive) words together when they correspond to a single word in the opposite language
- $\Rightarrow$ We can obtain a tokenisation suited to the case of bilingual word alignment

A bootstrap approach

- Word-packing can be performed using 1-to-n word alignment
- Word alignment can benefit from word packing

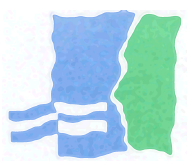


More at Yanjun's talk...

Recent addition (2): Super-Tagging

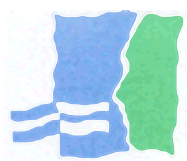
Integrating Syntax into SMT using supertags (Hassan et al., ACL 2006)

- Syntax modeled with supertags (CCG, TAG)
- one idea consists of giving preference to sequences of words that form a valid sequence of supertags
- useful for languages with different constituent orders $\Rightarrow$ syntax-driven re-ordering



Ongoing and Future Work

- Suffix-array based decoding (EBMT, phrase-based)
- Taking syntactical features into account in various contexts
- Synchronisation with work on tree-to-tree translations (more at Mary's talk)
- ...



Thank you

Thank you for your attention

http://www.nclt.dcu.ie/mt/