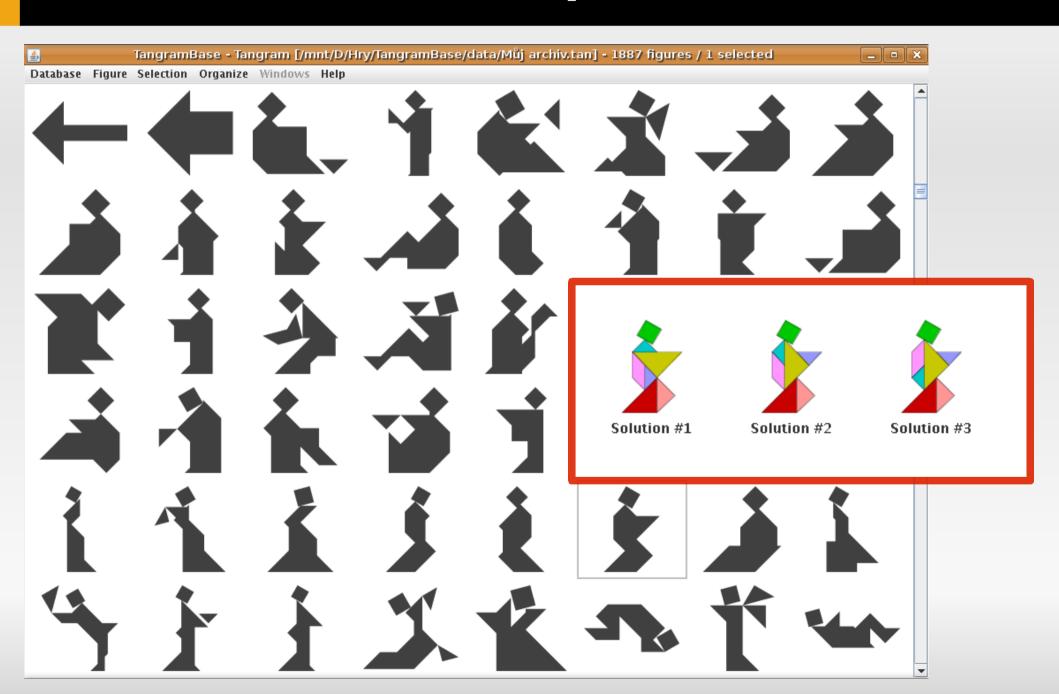
### 灰男孩

# Jan Popelka



### Background

- Bachelor degree at CTU in Prague
  - in Electrical Engineering and Informatics, Computer Science
  - abandoned Computer Graphics master program
- freelance and hobby experience with various programming languages
- interest in natural languages, backed up by frequent active use of Esperanto
- technical skills vs. scientific research?

### **Current Projects**

#### Annotation of coreference in PCEDT

- helped by and collaborating with MM, JŠ, ZŽ
- automatic grammatical coreference
- TrEd extension maintanance
- data distribution and technical support to annotators

### Diploma thesis supervised by Pavel Pecina

 Automatic acquisition of translation dictionaries from parallel corpora

- Wider view
  - Purpose
    - human-readable
    - computer-readable
    - MT only
  - Data source
    - parallel corpora × comparable corpora
    - plain text × information rich (i.e. Annotated)
  - Method
    - supervised × unsupervised

- Methods; things to consider
  - domain specificity
  - language dependency
  - coverage; recall and accuracy trade-off
  - time and memory complexity for large corpora
- Output
  - plain word-pairs, probabilistic dictionary, translation confidence
  - evaluation method (AER, BLEU)

### The thesis, guideliness

Parallel corpora, being the main source of training data for MT systems, can also be used for a simpler task — automatic acquisition of translation dictionary. The goal is to provide possible translational equivalents (in the target language) for each word of the source language, based on transcooccurrence statistics collected in a corpus.

w-alignment × t-alignment ?

#### pcedt rd 00393-s13 Věděl to nic také výjimečného není He 0.2792 1.005 also knew 0.38640.3650 0.2467 0.5674 he 0.9676 was n't 0.3897unique

#### **Our Approach**

- discriminative method
  - × generative, noisy channel (e.g. GIZA++)
- model combining various association measures
   [Pavel Pecina 2006] and (would-be) linguistic features
- combinatoric algorithm for finding the optimal alignment ( × incomplete sub-optimal search)
  - maximum weight edge cover [Jana Kravalová, 2007]
- feature engineering

- Some history and names
  - 1994 Dekai Wu and Xuanyin Xia
  - 1996 Melamed
  - 2005 Moore
  - 2005 Taskar and Lacoste-Julien
  - 2006 Blunsom and Cohn
  - 2008 Niehues and Vogel
  - 2008 Wei Chen
  - 2009 Yang Liu et al.

- discouraging results so far
  - GIZA++ baseline for intermidiate WA

$$AER = 0.186 P = 0.737 R = 0.926$$

Our best

```
AER = 0.246 P = 0.705 R = 0.820
/// SumSquaredError, QuasiNewton, 4:1, net 7:6:2
```

- still many things to try
  - more transparent model to see feature weights
  - collocations have to be addressed specifically

The Framework (C++, extensibility, reusability?)

- Features
  - using AM and trans-coocurrence statistics
  - not using AM, usually linguistically motivated
- generic features
  - parameters ~ idea
- derived features
  - base feature(s), derived bigram (parent, preceding)
- feature 'arity' (bigram/unigram only, row/col ...)

谢谢

Thank you