Document-Level Decoding in Moses

Nicola Bertoldi
Robert Grabowski
Liane Guillou
Michal Novak
Sorin Slavescu
Jose de Souza

MT Marathon 2012

Saturday 8th September, 2012
Problem of Lexical Consistency

Problem: inconsistent translation of words / phrases within a document

**Input**
con la possibilità di qualche pioggia
qualche pioggia

**Output**
chance of weak precipitations .
rainfalls

**Desired Output**
chance of weak precipitations .
some precipitation
Cache-Based Model

- Translate sentence by sentence using a Moses phrase-based system
- Cache unigrams from best translation of previously translated sentences
- At decoding, if word is found in the cache, add reward
- Designed as a generic framework:
  - Can be extended to the phrase-level
  - Can be used for topic and translation models
- Implemented LM Cache
- Method selected for simplicity - given time constraints
Wrapper

- Manages the translation process
- Takes unigrams from the sentence translation and formats XML input for the LM Cache

Possible extensions:
- Filtering to include only content words
- Using confidence estimation
Interface

- Between LM Cache / Scorer and Wrapper
- XML format: vector of strings from target translation
- E.g. Same weight: `<dlt trg="rain||rainfalls"/> pioggia`
- E.g. Different weights: `<dlt trg="rain"/> <dlt trg="rainfalls"/> pioggia`
LM Cache and Scorer

- **Cache:**
  - Stores unigrams with an age and score
  - Score decays exponentially as age increases
  - Can be initialised from a file e.g. with topic words etc.
  - Filled dynamically with unigrams from translation output of previous sentences or post-edits

- **Scorer:**
  - Checks cache for presence of words
  - Computes score for current sentence
  - Integrated as a feature function in the log linear model
Limitations

- Not yet thread-safe
- LM Cache only contains unigrams
- No filtering of unigrams
- Risk of error propagation - can be mitigated by a static load to initialise the LM Cache
Experimental Setup

- Italian -> English weather reports
- Data is repetitive and contains ambiguous words
- Test 3 scenarios in which lexical choice may be influenced:
  - Standard baseline
  - Using an inline suggestion - could be from previous sentence translations, etc.
  - Start with initialised LM Cache - from file
Scenario 1

**Scenario:** Standard baseline

**Input:** possibilità di qualche pioggia

- chance of weak precipitations . (-15.0119)
- chance of rainfalls . (-16.0039)
- chance of weak rainfalls . (-16.4628)
- the chance of rainfalls . (-16.6108)
- chance of rain . (-16.9722)
- rainfalls are likely . (-17.1521)
- the chance of some rain . (-17.4136)
- light rainfalls . (-17.4729)
- a chance of rainfalls . (-17.564)
- likelihood of rainfalls . (-17.5873)
Scenario 2

**Scenario:** Using an inline suggestion

**Input:** `<dlt -trg="rainfalls"/>` con la possibilità di qualche pioggia

- chance of rainfalls. (-14.6448)
- chance of weak precipitations. (-15.0119)
- chance of weak rainfalls. (-15.1037)
- the chance of rainfalls. (-15.2517)
- rainfalls are likely. (-15.7929)
- light rainfalls. (-16.1138)
- a chance of rainfalls. (-16.2048)
- likelihood of rainfalls. (-16.2281)
- some rainfalls. (-16.3367)
- possible rainfalls. (-16.4999)
Scenario 3

**Scenario:** Start with initialised LM Cache - file contains suggestion "rainfalls"

**Input:** possibilità di qualche pioggia

- chance of rainfalls . (-14.6448)
- chance of weak precipitations . (-15.0119)
- chance of weak rainfalls . (-15.1037)
- the chance of rainfalls . (-15.2517)
- rainfalls are likely . (-15.7929)
- light rainfalls . (-16.1138)
- a chance of rainfalls . (-16.2048)
- likelihood of rainfalls . (-16.2281)
- some rainfalls . (-16.3367)
- possible rainfalls . (-16.4999)
Re-cap (1)

- Designed and developed a generic framework for document-level decoding
- Applied framework to problem of lexical consistency
- Implemented a LM Cache for unigrams
- Wrapper used to provide suggestions to influence translation
Re-cap (2)

N-Best of the baseline for "qualche pioggia"

rainfalls (-14.5073)
rain (-14.5336)
some rain (-16.434)
weak rain (-16.8874)
some precipitation (-16.9085)

N-Best of the wrapper

rainfalls (-14.5073)
rain (-14.5336)
some precipitation (-15.5494)
some rain (-16.434)
light precipitation (-16.5994)
Re-cap (3)

- Possible extensions to Translation Models, Topic Models, n-grams, filtering
- Lots left to do...
Code Available

https://github.com/moses-smt/mosesdecoder/tree/moses_cachebased