Towards a General and Extensible Phrase-Extraction Algorithm

Wang Ling, L²F, INESC-ID/IST
Tiago Luís, L²F, INESC-ID/IST
João Graça, L²F, INESC-ID/IST
Luísa Coheur, L²F, INESC-ID/IST
Isabel Trancoso, L²F, INESC-ID/IST

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa
Laboratório de Sistemas de Língua Falada

{wang.ling,tiago.luis,joao.graca,luisa.coheur,isabel.trancoso}@l2f.inesc-id.pt

Motivation

Phrase Extraction

- Plays a key role in enriching translation quality
- However:
  - ♢ there is not a general algorithm that allows this process to be easily extended
  - ♢ a great deal of time is spent reimplementing the same generic algorithm

Goals

- Create a general algorithm for phrase extraction that is easily entensible
- Identify key control points that define the behavior of the algorithm
- Provide modular interface to allow users to extend these control points

1

Algorithm - Extract Phrase Pairs

```
Require: Bilingual sentence s
  fl = s.foreignLen
  sl = s.sourceLen
  extractedPhrasePairs = {}
  for fp = 0; fp ≤ fl; fp++ do
    for fd = 1; fd ≤ maxDuration; fd++ do
      if ForeignPhraseAcceptor.accept(s, fp, fd) then
        for sp = 0; sp ≤ sl; sp++ do
          for sd = 1; sd ≤ maxDuration; sd++ do
            if SourcePhraseAcceptor.accept(s, sp, sd) then
              PhrasePair p = phrase pair from s from (fp, sp) to (fd, sd)
              LocalPhrasePairFeaturesCreator.addFeatures(p)
              if PhrasePairAcceptor.accept(p) then
                extractedPhrasePairs.add(p)
              end if
            end if
          end for
        end for
      end if
    end for
  end for
  return extractedPhrasePairs
```

Extract Phrase Pairs

- Extract all phrase pairs from a given sentence pair
- Invoked in the context of the General Phrase Extraction

Key Control Points

- SourcePhraseAcceptor - Decide if the source phrase is a good translation unit
- ForeignPhraseAcceptor - Decide if the foreign phrase is a good translation unit
- LocalPhrasePairFeaturesCreator - Calculates the features for the extracted phrase pair
- PhrasePairAcceptor - Decides whether the pair phrase is a suitable candidate for extraction

2

Algorithm - General Phrase Extraction

```
Require: Bilingual Corpus
Require: MaximumPhraseSize - max
  for each sentence pair (s, t) in Corpus do
    extractedPhrasePairs = extractPhrasePairs(s, t, max)
    for each phrase pair p in extractedPhrasePairs do
      phraseTable.add(p)
    end for
  end for
  computeGlobalPhraseStats
  pruneGlobalPhraseStats
  savePhraseTable
```

General Phrase Extraction

- Extract all phrase pairs from a given bilingual corpora

Key Control Points

- ComputeGlobalPhraseStats - Merges extracted pairs and calculates their features
- PruneGlobalPhraseStats - Eliminates phrase pairs

3

Implemented Extensions

Moses pipeline / Baseline

- SizeBasedSourceAcceptor and SizeBasedTargetAcceptor
- KohenAcceptor
- ProbabilityFeatureCreator
- LexicalWeightingFeatureCreator

Weighted Alignments

- WeightedAlignmentScorer
- WeightedAlignmentAcceptor

Punctuation Based Filtering

- NoPunctuationAcceptor
- NoTerminalPunctuationAcceptor

4

Experiments

Method	BTEC (Fr-En)	DIALOG (Cn-En)
Baseline / Moses	62.46	41.42
Weighted Alignments	63.07	42.15
NoPunct	62.75	41.20
NoTerminalPunct	**63.41**	**42.28**

- Using weighted alignments and NoTerminalPunct accepter improved the results

5

Conclusions

- Introduced a general algorithm for phrase extraction
- Easily extensible by adding implementations of control points
- Available at *https://code.google.com/p/geppetto*

6

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007. The PhD thesis of Tiago Luís is supported by FCT grant SFRH/BD/62151/2009. The PhD thesis of Wang Ling is supported by FCT grant SFRH/BD/51157/2010.