Spoken Language Translation through Confusion Network decoding

Nicola Bertoldi
FBK-irst, Trento, Italy

Edinburgh, 20 April 2007
Outline

• Spoken Language Translation
  – task
  – specific issues
  – formal definition
  – common approaches

• SLT by Confusion Network decoding
  – definition of Confusion Network
  – CN decoding algorithm
  – efficiency
  – advanced features of Moses and CN
  – evaluation

• Other applications of CN decoding

Credits: R. Zens (RWTH, Aachen), M. Federico (FBK-irst, Trento)
Spoken Language Translation

- **Translation from speech input**
  - recent and challenging task of Machine Translation

- **Combination of ASR and MT:**
  - *cascade* of ASR and MT systems
  - different *interfaces*, different approaches

- **Harder** than text translation
  - input genre is more *spontaneous*
  - ASR is far from being a solved problem
    - *transcription errors* are generated
    - *punctuation* is missing (or post-added)
    - *case information* is (often) missing
SLT issues

"and ... then ... here we have seen success"

Speech Signal:

Correct Transcription: and @ehm then @mh here we have seen success

Best ASR Transcription: and me @mh there we have seen a success

- transcription errors: substitution, insertion, deletion
- spontaneous speech phenomena: hesitation, repetition
SLT issues

- spontaneous speech phenomena can cause
  - *transcription errors*:
    and *@ehm then here* we have seen $\rightarrow$ and *me there* we have seen
    *@uh I see* $\rightarrow$ *you see*
  - *bad-formed* sentence
    *mister mister @ehm mister maaten*

- transcription errors modify both *meaning* and *syntax*:
  - *semantic errors*:
    *mister maaten* has the floor $\rightarrow$ *mister martin* has the floor
    *market* $\rightarrow$ *mark at* *ate* $\rightarrow$ *eight* *you* $\rightarrow$ *e.u.*
  - *syntactic errors*:
    *I move on to the committee* $\rightarrow$ *I'll move onto the committee*
    *@uh I see* $\rightarrow$ *you see*
SLT issues

• transcription and translation quality *strongly correlate*  
  – the better transcription, the better translation

• ASR quality increases in a set of transcription hypotheses

• but unfortunately the *oracle* is unknown

⇒ *translation of as many alternative transcriptions* as possible

• In principle:  
  – all transcriptions in the *Word Graph* generated by the ASR system
Word Graph

- large amount of transcription hyps produced by the ASR system
- arcs are labelled with words and ASR scores
- nodes are labelled with starting and ending times of words
- **redundancy** is high (from the point of view of MT):
  - many paths represent the same hyp differing just in timestamps
- topology is **complex** (from the point of view of MT):
  - word-coverage and word-reordering are hard to handle
Approaches to SLT

- different *approximations* of a WG
- different *interfaces*:
  - 1-best, *N*-best, *confusion network*
  - full word graph
- *dedicated* MT decoder

- Finite State Transducer:
  - ASR and MT models merged into one finite-state network
  - a transducer decodes the input speech in one shot
  - difficult scaling up to very large domains

- [Casacuberta et al., CSL, 2004]
Statistical Spoken Language Translation

Given a *speech input* \( o \) in the source language, find the *best translation* through the following approximate criterion:

\[
e^* = \arg \max_e \Pr(e \mid o) = \arg \max_e \sum_{f \in F(o)} \Pr(e, f \mid o)
\]

\[
\approx \arg \max_e \max_{f \in F(o)} \Pr(e, f \mid o)
\]

- \( F(o) \) is any *set of possible transcriptions* of \( o \)
  - interface between ASR and MT
- \( \Pr(e, f \mid o) \) is any *phrase-based speech translation model*
- the actual transcription \( f \) is regarded as a hidden variable
- approximation simplifies the search algorithm
1-best Decoder

- translation of the *first best* transcription only
- use of a *standard MT system* of text

- no multiple transcriptions
- impossible recover from ASR errors
$N$-best Decoder

- translation of $N$-best transcription hypotheses
- **rerank** with additional ASR scores
  - acoustic likelihood and source LM

1. and there we have seen a success -217 -12
2. and there we have seen success -198 -9

8. and then here we have seen success -215 -21
9. and now here we have seen a success -265 -3

- possible recover from ASR errors
- no exploitation of overlaps among $N$-best
Confusion Network Decoder

- translation of a confusion network, a compact structure approximating a WG
- exploitation of multiple transcription hypotheses
- exploitation of overlaps among hypotheses
- extension of a standard text decoder

[ASRU, 2005], [ICASSP, 2007], Moses’ doc
A **Confusion Network** approximates a WG by a linear network, s.t.: 

- arcs are labeled with words or with the *empty word* ($\epsilon$-word) 
- arcs are weighted with word *posterior probabilities* 
- paths are a superset of those in the word graph 
- paths can have different lengths
Extraction of CN from WG

- *cluster nodes* with close timestamps
- possibly *introduce special arcs* for empty-words
- *compute word posterior probabilities* exploiting ASR scores

Statistical model for CN decoding

- Translation Model is a log-linear combination of features
- Features are defined in terms of phrases
- Standard feature functions for text decoder:
  - Language Models
  - Distortion Model
  - Lexicon Model (LexM)
  - Phrase and Word Penalties

- Specific feature functions for Confusion Network (CM)
  - likelihood of the path into the source CN: product of word posterior probs
  - number of words in the path (optional)

- LexM and CM depend on the source phrase:
  - different paths in a span give different scores
Translation from text

- **cover** a not yet covered span
  - one source phrase

- **retrieve** all translation options
  - looking up into the phrase table

- **compute** feature scores
- **recombine** hypotheses
- ...

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>cannot</td>
<td>say</td>
<td>anything</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cannot say</th>
</tr>
</thead>
<tbody>
<tr>
<td>no puedo decir</td>
</tr>
<tr>
<td>ella no puede decir</td>
</tr>
<tr>
<td>él no puede decir</td>
</tr>
<tr>
<td>.....</td>
</tr>
</tbody>
</table>
Translation from Confusion Network

Extension of the translation from text

- **cover** a not yet covered *span*
  - *many source phrases*

- **retrieve** all translation options
  - for all source phrases in the span
  - looking up into the phrase table

- **compute** scores
- **recombine** hypotheses
- ...

<table>
<thead>
<tr>
<th>I</th>
<th>cannot</th>
<th><em>eps</em></th>
<th>say</th>
<th><em>eps</em></th>
<th>anything</th>
</tr>
</thead>
<tbody>
<tr>
<td>hi</td>
<td>can</td>
<td>not</td>
<td>said</td>
<td>any</td>
<td>thing</td>
</tr>
<tr>
<td></td>
<td><em>eps</em></td>
<td></td>
<td>says</td>
<td>things</td>
<td></td>
</tr>
</tbody>
</table>

| cannot say
| cannot said
| can say
| not say
| ....

| no puedo decir
| no puede decir
| ella no puede decir
| él no puede decir
| ....
| ........

| puedo decir
| puede decir
| ella puede decir
| él puede decir
| ....
| ........

phrase table

phrase table
Issues of CN Decoding

- Number of paths grows \textit{exponentially} with span length
- Look-up of translations for a huge number of source phrases
- \textit{Enumeration} of all alternatives is \textit{unfeasible}
- and \textit{dummy}!

Indeed:

- Paths can correspond to phrases without translations

\[
\begin{align*}
those_{0.92} & & \epsilon_{0.07} \\
& as_{6e-4} & \text{is}_{1e-5} \\
& there_{5e-5} & \text{who}_{1e-5} \\
& \text{who's}_{5e-6} & \text{were}_{0.99} \\
wel_{7e-6} & \epsilon_{1e-5} \\
\text{was}_{8e-6} & \epsilon_{1e-5}
\end{align*}
\]
Issues of CN Decoding

• different paths into a span can correspond to the same phrase (who was)
  – different CM score

  \[
  \begin{array}{cccc}
  \text{those} & \epsilon_{0.92} & \epsilon_{0.99} & \epsilon_{0.99} \\
  \text{was} & _{6e-5} & \text{were} & \text{well} \\
  \text{as} & _{6e-4} & \text{is} & _{1e-5} \\
  \text{there} & _{5e-5} & \text{who} & _{2e-6} \\
  \text{who's} & _{5e-6} & & \\
  \end{array}
  \]

• different phrases into the same span can have equal translation
  – who’s who and who is who translates into quién es quién
  – different CM and LexM scores

  \[
  \begin{array}{cccc}
  \text{those} & \epsilon & \epsilon & \epsilon \\
  \text{was} & \text{well} & \text{as} & \text{is} \\
  \text{there} & \text{who} & \text{who} & \text{who} \\
  \text{who's} & & & \\
  \end{array}
  \]
Solution for an efficient CN decoding

- **Optimization of the retrieval of the translation options** by:
  - representing source entries of the phrase-table as *prefix-trees*
  - *incrementally pre-fetching* translation options
  - *early recombining* translation options

- **Once translation options are generated, usual decoding applies.**
Prefix-tree representation of phrase table

source phrases

1 word

2 words

3 words

target phrases

translations

.....

puedo

puede

ella puede

.....

no puedo decir

él puede no decir

ella no puede decir

él no puede decir

.....

no puedo

puede no

ella no puede

él no puede

.....

N. Bertoldi

SLT through CN decoding

Edinburgh, 20 April 2007
Incremental pre-fetching of translation options

- collect translation options *incrementally over the span length*
  - exploit knowledge about shorter span

- *once* and *before decoding*

- **worst case** (all phrases are present) is still exponential, but *never happens*
Early recombination

- *Different phrases* into the same span can have the *same translation*
- *Different LexM* and *CM* scores, the other are equal
- *Undistinguishable* from the decoder

- Take the *best path* only (and its scores)

- Use $LexM(span, e)$ and $CM(span)$, instead of $LexM(f, e)$ and $CM(f)$

\[
\begin{align*}
LexM(span, e) &= LexM(\hat{f}, e) \\
CM(span, e) &= CM(\hat{f}, e) \\
\hat{f} &= \arg \max_{f \in \text{span}} \lambda_{LexM}LexM(f, e) + \lambda_{CM}CM(f)
\end{align*}
\]
Efficiency of Search Algorithm

N. Bertoldi

SLT through CN decoding

Edinburgh, 20 April 2007
CN decoding in Moses

- Moses implements CN decoding

- **Factored models**
  - alternative over the full factor space

  \[
  \begin{array}{lll}
  \text{Haus|N} & \text{der|ART} & \text{Zeitung|N} \\
  \text{aus|PREP} & \text{des|ART} & \epsilon|\epsilon \\
  \text{aus|ADV} & \epsilon|\epsilon & \text{Zeitungs|N} \\
  \epsilon|\epsilon & \text{drei|N} & \text{Zeitungen|N}
  \end{array}
  \]

- **Lexicalized Distortion Models**
  - conditioned on the best path inside a span
### CN decoding: results

- Spanish-English EPPS 2006 Evaluation

<table>
<thead>
<tr>
<th>type</th>
<th>WER</th>
<th>BLEU</th>
<th>NIST</th>
<th>PER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>verbatim</td>
<td>0.0</td>
<td>48.00</td>
<td>9.864</td>
<td>31.19</td>
<td>40.96</td>
</tr>
<tr>
<td>cn-oracle</td>
<td>8.45</td>
<td>44.12</td>
<td>9.356</td>
<td>34.37</td>
<td>44.95</td>
</tr>
<tr>
<td>cons-dec</td>
<td>23.30</td>
<td>36.98</td>
<td>8.550</td>
<td>39.17</td>
<td>49.98</td>
</tr>
<tr>
<td>cn</td>
<td>8.45</td>
<td>39.17</td>
<td>8.716</td>
<td>38.64</td>
<td>49.52</td>
</tr>
<tr>
<td>1-best</td>
<td>22.41</td>
<td>37.57</td>
<td>8.590</td>
<td>39.24</td>
<td>50.01</td>
</tr>
<tr>
<td>5-best</td>
<td>18.61</td>
<td>38.68</td>
<td>8.694</td>
<td>38.55</td>
<td>49.33</td>
</tr>
<tr>
<td>10-best</td>
<td>17.12</td>
<td>38.61</td>
<td>8.694</td>
<td>38.69</td>
<td>49.46</td>
</tr>
</tbody>
</table>

- Relative Improvement in BLEU: 30% (wrt to oracle)
- CN decoding speed is 2 times slower
CN decoding: results

- Moses vs. Irst-05 vs. Irst-06

<table>
<thead>
<tr>
<th>Input</th>
<th>WER</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>verbatim</td>
<td>0.0</td>
<td>40.84</td>
</tr>
<tr>
<td>1-best</td>
<td>14.61</td>
<td>36.64</td>
</tr>
<tr>
<td>cons-dec</td>
<td>14.46</td>
<td>36.54</td>
</tr>
<tr>
<td>cn</td>
<td>11.61</td>
<td>37.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Irst-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Irst-06 was top system
- Irst-05 and Irst-06 translate pruned confusion networks
- Irst-05 translates CN 18 times slower than text
Other applications of CN decoder

- **CN represents ambiguity**
  - variations, alternatives, errors

- CN decoder *disambiguates* and *translates* in one shot:
  - *insertion of punctuation* and case restoring in translation

- CN decoder is also a *tagger*:
  - POS tagging, case restoring
  - Word Sense Disambiguation, NE Recognition, OCR, etc.
  - using monotone translation
  - using ad-hoc lexicon models and LMs

<table>
<thead>
<tr>
<th>I@P</th>
<th>read@VP</th>
<th>a@R</th>
<th>book@N</th>
<th>thank</th>
<th>you</th>
<th>mr.</th>
<th>bond</th>
</tr>
</thead>
<tbody>
<tr>
<td>read@VPP</td>
<td>book@VP</td>
<td>book@VI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>read@VI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thank You Mr. Bond
Punctuating Confusion Networks

Confusion network without punctuation

Consensus decoding

Punctuating confusion network

Punctuated confusion network
Punctuating Confusion Networks: Results

- ASR 1-best output vs. confusion network
- 1-best punctuation vs. punctuating CN (from 1K-best)

### Spanish-English EPPS Eval06

<table>
<thead>
<tr>
<th>ASR type</th>
<th>punctuation</th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-best</td>
<td>1-best</td>
<td>35.62</td>
<td>8.37</td>
<td>57.15</td>
<td>44.56</td>
</tr>
<tr>
<td></td>
<td>CN</td>
<td>36.01</td>
<td>8.41</td>
<td>56.78</td>
<td>44.39</td>
</tr>
<tr>
<td>CN</td>
<td>1-best</td>
<td>36.22</td>
<td>8.46</td>
<td>56.39</td>
<td>44.37</td>
</tr>
<tr>
<td></td>
<td>CN</td>
<td>36.45</td>
<td>8.49</td>
<td>56.17</td>
<td>44.19</td>
</tr>
</tbody>
</table>
Conclusion

- Spoken Language Translation
- SLT system:
  - combination of ASR and MT through Confusion Network
  - effective representation of a huge number of transcription hypotheses
- Efficient search algorithm for CN-based SMT:
  - prefix-tree representation and pre-fetching of lexicon models
  - early recombination of translation options
- Moses system:
  - CN decoding
  - state-of-the-art for SLT (translation performance and decoding speed)
  - slight improvement of CN decoder vs. 1-best decoder
- Moses for enriched translation
- Moses for tagging
Thank you!