Comparative Evaluation of Research vs. Online MT Systems

Antonio Toral†, Federico Gaspari†, Sudip Kumar Naskar†* and Andy Way†*

CoSyne† / CNGL*

School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland

{atoral, fgaspari, snaskar, away}@computing.dcu.ie
Outline

• Introduction

• MT evaluation metrics

• Data sets

• MT systems

• Results

• Conclusion

• Future work
Introduction

- Experiments at the end of Y1 of 3-year EU-funded CoSyne project
- Three language pairs evaluated on data from the news domain
  - DE → EN
  - IT → EN
  - NL → EN
- Compared CoSyne MT system against 4 free web-based systems
- Wide range of state-of-the-art automatic evaluation metrics used
Introduction: CoSyne

- FP7 STREP project (call 4, objective 2.2)
- 3 years: Mar 2010 – Feb 2013

Objective: to automate the dynamic multilingual content synchronization process of wikis across languages

Languages
- 4 core languages: English, German, Italian and Dutch
- 2 less resourced languages: Turkish and Bulgarian (year 3)

The CoSyne system will be integrated via web services with the open-source MediaWiki package

The overall CoSyne system includes
- Document structure modeling
- Document structure induction
- Textual entailment
- Machine translation
Introduction: CoSyne

• Consortium
  ○ 7 partners from 4 EU countries: Germany, Ireland, Italy and the Netherlands

• 3 academic partners
  ○ University of Amsterdam (UvA)
  ○ Fondazione Bruno Kessler (FBK)
  ○ Dublin City University (DCU)

• 1 research organization
  ○ Heidelberg Institute for Theoretical Studies (HITS)

• 3 end users
  ○ Deutsche Welle (DW) - Poster in EU projects section!
  ○ Netherlands Institute for Sound and Vision (NISV)
  ○ Vereniging Wikimedia Nederland (VWN)
MT evaluation metrics

• BLEU (Papineni et al., 2002)
  \[ n\text{-gram based} \]

• NIST (Doddington, 2002)

• GTM (Turian et al., 2003) \textit{based on standard measures in NLP}

• METEOR (Banerjee and Lavie, 2005)
  ° METEOR-NEXT \textit{additional linguistic information (stemming, synonyms)}

• TER (Snover et al., 2006)
  ° TERp \textit{error rates}

• DCU-LFG (Owczarzak et al., 2007; He et al., 2010) \textit{syntactic dependencies}
Data sets

• From the news domain
  ◦ to match usage scenarios envisaged by end users

• Language pairs
  ◦ DE — EN
  ◦ IT — EN
  ◦ NL — EN

• 2,000 sentence pairs per language combination
  ◦ 1k development + 1k evaluation
Data for DE—EN (DW)

• Documents provided by DW from two online journals
  ◦ Europa Aktuell 2001 to 2010: 2,201 documents
  ◦ Global 3,000: 80 documents

• XML format + alignment scores (sentence, doc, etc.)
  ◦ DE: 25,797 words (average sentence length: 12.9 words)
  ◦ EN: 26,938 words (average sentence length: 13.47 words)

• Tools used to align text
  ◦ TreeTagger
  ◦ Hunalign along with a bilingual dictionary derived from Apertium’s DE—EN dictionary
Data for IT—EN (DCU)

• Manual download and alignment of parallel documents
• AsiaNews website (up to July 2010): 87 document pairs
  ◦ IT: 38,607 words (average document length: 444 words)
  ◦ EN: 38,090 words (average document length: 438 words)
Data for NL—EN (NISV)

• Three different data sets:
  ◦ België Diplomatie: 418 HTML doc pairs
  ◦ Video Active: 1,076 doc pairs (XML format)
  ◦ NISV Wiki: 30 doc pairs
    • NL: 45,546 words (average sentence length: 22.8 words)
    • EN: 46,390 words (average sentence length: 23.2 words)

• Same tools used to align text as for DE—EN
  ◦ Bilingual dictionary from Apertium’s NL—EN dictionary
MT systems

• Free online MT systems
  ◦ Statistical systems
    • Google Translate (Google)
    • Bing Translator (Microsoft)
  ◦ Rule-based systems
    • Systran
    • FreeTranslation (SDL)

• The research MT system
  ◦ CoSyne MT system (Martzoukos & Monz, 2010)
    • Statistical system
    • Developed by UvA (thanks to Christof Monz and his team for making it available for these experiments)
Recap – MT evaluation results

- Language pairs
  - DE → EN
  - IT → EN
  - NL → EN

- MT systems
  - Google Translate
  - Bing Translator
  - Systran
  - FreeTranslation
  - CoSyne MT system at M12

- MT evaluation metrics
  - BLEU, METEOR, METEOR-NEXT, GTM, DCU-LFG, NIST, TER and TERp
  - Statistical significance tests provided only for BLEU, NIST and GTM

Divided by a factor of 10 for consistency

Given as 1-x to reverse the trend for comparability
## DE → EN results

| de-en       | BLEU  | NIST   | METEOR | METEOR-NEXT | TERp  | TER   | GTM    | DCU-LFG |
|-------------|-------|--------|--------|-------------|-------|-------|--------|---------|
| **BLEU**    | 0.2477_{b,c,d,e} | 0.2294_{c,d} | 0.1752_{d} | 0.1657 | 0.2052_{c,d} |
| **NIST**    | 0.6358_{c,d,e} | 0.6362_{c,d,e} | 0.5447_{d} | 0.5212 | 0.5788_{c,d} |
| **METEOR**  | 0.5830 | 0.5584 | 0.5239 | 0.5060 | 0.5470 |
| **METEOR-NEXT** | 0.4977 | 0.4807 | 0.4552 | 0.4422 | 0.4692 |
| **TERp**    | 0.4000 | 0.3600 | 0.3216 | 0.3100 | 0.2941 |
| **TER**     | 0.4172 | 0.4161 | 0.3444 | 0.3273 | 0.3700 |
| **GTM**     | 0.4517_{b,c,d,e} | 0.4270_{c,d,e} | 0.4057_{d,e} | 0.3849 | 0.3914 |
| **DCU-LFG** | 0.4899 | 0.4570 | 0.4133 | 0.3957 | 0.4261 |
• SMT systems perform better than RBMT systems

• For most metrics (except TERp and GTM) the performance of the CoSyne MT system is between SMT systems and RBMT systems

• Google beats Bing according to almost all metrics (NIST is a tie)

• Systran outperforms FreeTranslation across all the metrics
IT → EN results

| Measure          | Google | Bing   | Systran | Freetranslation | CoSyne M12 |
|------------------|--------|--------|---------|-----------------|------------|
| BLEU             | 0.4235 | 0.3106 | 0.1840  | 0.1754          | 0.3137     |
| NIST             | 0.8579 | 0.7517 | 0.5439  | 0.5427          | 0.7318     |
| METEOR           | 0.7017 | 0.6384 | 0.5709  | 0.5537          | 0.6565     |
| METEOR-NEXT      | 0.5942 | 0.5412 | 0.4832  | 0.4700          | 0.5545     |
| TERp             | 0.5600 | 0.4700 | 0.3890  | 0.3800          | 0.4946     |
| TER              | 0.5599 | 0.4857 | 0.3225  | 0.3128          | 0.4679     |
| GTM              | 0.6187 | 0.5394 | 0.4596  | 0.4510          | 0.5475     |
| DCU-LFG          | 0.6400 | 0.5200 | 0.4244  | 0.4080          | 0.5311     |
IT → EN discussion

• Google is the clear top performer across all the metrics

• CoSyne MT system performs better than Bing in most metrics (except NIST and TER, but close scores)

• The two RBMT systems attain very similar scores for all evaluation metrics, showing much poorer performances than the SMT systems
### NL → EN results

| n1-en          | Google | Bing | Systran | Freetranslation | CoSyne M12 |
|----------------|--------|------|---------|----------------|------------|
| **BLEU**       | 0.3330 | 0.3347 | 0.2643 | 0.2456         | 0.3223     |
| **NIST**       | 0.7986 | 0.7596 | 0.6830 | 0.6479         | 0.7532     |
| **METEOR**     | 0.6633 | 0.6695 | 0.6161 | 0.5964         | 0.6431     |
| **METEOR-NEXT**| 0.5583 | 0.5628 | 0.5180 | 0.5032         | 0.5419     |
| **TERp**       | 0.4987 | 0.5066 | 0.4315 | 0.4123         | 0.4690     |
| **TER**        | 0.5251 | 0.4892 | 0.4424 | 0.4221         | 0.5000     |
| **GTM**        | 0.5339 | 0.5156 | 0.4761 | 0.4672         | 0.4956     |
| **DCU-LFG**    | 0.5459 | 0.5507 | 0.4661 | 0.4411         | 0.5080     |
SMT systems outperform RBMT systems.

Google outperforms Bing for NIST, TER and GTM, while for the other metrics Bing receives higher scores.

CoSyne MT system performs better than the RBMT systems and is close to Google and Bing.

- CoSyne MT system performs better than Bing according to TER.
Summary of results

- The three SMT systems receive (much) higher scores than the two RBMT systems for all the 8 evaluation metrics in each of the 3 language pairs.

- Overall Google Translate receives the best scores consistently across most of the metrics for all 3 language pairs.

- Bing Translator and the CoSyne MT system perform similarly:
  - Inferior than Google, but better than Systran and FreeTranslation.

- CoSyne good for IT / NL → EN, improvement needed for DE → EN.

- Among the RBMT systems, Systran always performs better than FreeTranslation according to all the 8 evaluation metrics for the 3 language pairs.
Conclusion

- CoSyne MT system evaluation (M12 implementation)
  - against four free online MT systems
  - for 3 language pairs (DE / IT / NL → EN)
  - across 8 automatic MT evaluation metrics

- Assessed the performance of the MT component of the CoSyne system against state-of-the-art MT systems

- Monitor progress over time

- Prioritize and focus efforts on the development and fine-tuning of language pairs requiring improvement (e.g. DE → EN)
Future work

• Diagnostic evaluation
  ◦ Analysis of TER results (INS, DEL, SUB, SHFT) [currently underway]
  ◦ Methodology based on linguistic checkpoints following Zhou et al. (2008) to evaluate system over any linguistic phenomenon [Feb. 2012]

• Correlations of automatic evaluation metrics with human judgments and perception of MT quality (staff at end user partners) [Aug. 2011]
Thank you for your attention!

Questions?

Comparative Evaluation of Research vs. Online MT Systems

Antonio Toral†, Federico Gaspari†, Sudip Kumar Naskar†* and Andy Way†*

CoSyne† / CNGL*
School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland

{ataloral, fgaspari, snaskar, away}@computing.dcu.ie