Automatic MT Evaluation
Overview

- Introduction
- Automatic MT evaluation
- Linguistically motivated evaluation measures
- Quality estimation
- The Asíya toolkit
MT Development cycle (1)

MT system developer

Evaluation Methods

Error Analysis
- Identify Type of Error
- Analyze Possible Causes

System Refinement

Evaluation
MT Development cycle (2)

Error Analysis
- Identify Type of Error
- Analyze Possible Causes
- System Refinement
- Evaluation

Evaluation Methods
- MT system developer
- MT metric developer
- Error Analysis
- Identify Type of Error
- Analyze Possible Causes
- Metric Refinement
- Evaluation

Meta-Evaluation Methods

MT Tutorial
Difficulties of the MT evaluation (1)

- Machine Translation is an open NLP problem
  - The correct translation is not unique.
  - The set of valid translations is not small.
  - The quality of a translation is a fuzzy concept.
Difficulties of the MT evaluation (2)

- Quality aspects are heterogeneous:
  - **Adequacy (or Fidelity):** Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?
  - **Fluency (or Intelligibility):** Is the output fluent? This involves both grammatical correctness and idiomatic word choices.
  - **Post-edition effort:** time required to repair the translation, number of key strokes, etc.
Example

- El mando de la Wii ayuda a diagnosticar una enfermedad ocular infantil.

- The remote control of the Wii helps to diagnose an infantile ocular disease.

- The control of the Wii help to diagnose an ocular illness childish.
- The control of the Wii helps to diagnose an infantile ocular disease.
- The Wii remote helps diagnose a childhood eye disease.
- The Wii Remote to help diagnose childhood eye disease.
- The control of the Wii helps to diagnose an ocular infantile disease.
- The mando of the Wii helps to diagnose an infantile ocular disease.
Manual vs. automatic evaluation

- Categorisation problem for human annotations.
  - 5-point likert scale [LDC05]
  - 4-point likert scale [TAUS13]

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>All</td>
<td>Flawless</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Most</td>
<td>Good</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Little</td>
<td>Disfluent</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>None</td>
<td>Incomprehensible</td>
</tr>
</tbody>
</table>

- Ranking problem for human annotations [Cal12]
- Regression problem for automatic metrics.
Meta-Evaluation

- Correlation with human assessments
  - Pearson (system level)
  - Spearman
  - Kendall’s tau (segment level)

- Consistency (ranking)

- AvgDelta [Cal12]
Human Annotation Tools

- BLAST [Sty11] - annotate
- Appraise [Fed12] - rank
- DQF [Tau12] – best practices
The head of the Turkish Ministry of Foreign Affairs Muammer Guler initially denied the construction, then announced that the wall is being constructed to provide border security. "The mayor of Nusaybin, Aisha Gokhan, a member of the BDP, went on a hunger strike, thus turning her protest against construction of the wall into a deadly fight," they announced in the party's press office.
Interannotator Agreement

- Cohen’s kappa coefficient [Coh60]
- WMT13 [Boj13]
- Kappa interpretation [Lan77]
  - 0.0–0.2 slight
  - 0.2–0.4 fair
  - 0.4–0.6 moderate
  - 0.6–0.8 substantial
  - 0.8–1.0 almost perfect

<table>
<thead>
<tr>
<th>Pair</th>
<th>Inter-κ</th>
<th>Intra-κ</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZ-EN</td>
<td>0.244</td>
<td>0.479</td>
</tr>
<tr>
<td>EN-CZ</td>
<td>0.168</td>
<td>0.290</td>
</tr>
<tr>
<td>DE-EN</td>
<td>0.299</td>
<td>0.535</td>
</tr>
<tr>
<td>EN-DE</td>
<td>0.267</td>
<td>0.498</td>
</tr>
<tr>
<td>ES-EN</td>
<td>0.277</td>
<td>0.575</td>
</tr>
<tr>
<td>EN-ES</td>
<td>0.206</td>
<td>0.492</td>
</tr>
<tr>
<td>FR-EN</td>
<td>0.275</td>
<td>0.578</td>
</tr>
<tr>
<td>EN-FR</td>
<td>0.231</td>
<td>0.495</td>
</tr>
<tr>
<td>RU-EN</td>
<td>0.278</td>
<td>0.450</td>
</tr>
<tr>
<td>EN-RU</td>
<td>0.243</td>
<td>0.513</td>
</tr>
</tbody>
</table>
Benefits of Automatic Evaluation (1)

- Compared to manual evaluation, automatic measures are:
  - Cheap (vs. costly)
  - Objective (vs. subjective)
  - Reusable (vs. not-reusable)
Benefits of Automatic Evaluation (2)

- Automatic evaluation metrics have notably accelerated the development cycle of MT systems
  - Error analysis
    - Identify and analyze weak points
  - System optimization
    - Ranking of N-best list and parameter estimation
  - System comparison
    - Phrase- or system-based combination
Active Topic of Research

- Annual metrics competition organized by the WMT workshop series and supported by the EC
  - http://www.statmt.org/wmt14/
  - Both Evaluation Measure and Confidence Estimation

- Biannual OpenMT metric competition organized by NIST and supported by DARPA
  - Evaluation Measures for informal data genres and speech translations

- 1st Workshop on Asian Translation, Tokyo, October 2014
  - http://orchid.kuee.kyoto-u.ac.jp/WAT/
  - Japanese-Chinese, test data is prepared using paragraph as a unit
Overview

- Introduction
- **Automatic MT evaluation**
- Linguistically motivated evaluation measures
- Quality estimation
- The Asiya toolkit
MT Automatic Evaluation (1)

- Setting:
  - Compute the similarity between a system's output and one or several reference translations.

- Challenge:
  - The similarity measure should be able to discriminate whether the two sentences convey the same meaning (semantic equivalence).
MT Automatic Evaluation (2)

- Goals:
  - Low cost
  - Tunable
  - Meaningful
  - Coherent
  - Consistent
First Approaches

- Lexical similarity as a measure of quality
  - Edit Distance: WER [Nie00], PER [Til97], TER [Sno06]
  - Precision: BLEU [Pap01], NIST [Dod02]
  - Recall: ROUGE [Lin04a]
  - Precision/Recall: GTM [Melo3], METEOR [Ban05,Den10]
The remote control of the Wii helps to diagnose an infantile ocular disease.

The Wii remote helps diagnose a childhood eye disease.
The remote control of the Wii helps to diagnose an infantile ocular disease.

The Wii remote helps diagnose a childhood eye disease.

- **Precision**: \[ \frac{\text{correct}}{\text{output\_length}} = \frac{7}{10} = 0.7 \]

- **Recall**: \[ \frac{\text{correct}}{\text{reference\_length}} = \frac{7}{14} = 0.5 \]

- **F-measure**: \[ \frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})/2} = \frac{0.35}{0.6} = 0.583 \]
The remote control of the Wii helps to diagnose an infantile ocular disease.

Wii the control of the remote to diagnose disease helps an ocular infantile.

- **Precision:** \[
\frac{\text{correct}}{\text{output\_length}} = \frac{14}{14} = 1.00
\]

- **Recall:** \[
\frac{\text{correct}}{\text{reference\_length}} = \frac{14}{14} = 1.00
\]

- **F-measure:** \[
\frac{\text{precision\_recall}}{(\text{precision} + \text{recall})/2} = \frac{1.00}{1.00} = 1.00
\]
“The main idea is to use a **weighted average of variable length phrase matches** against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family.” [Pap01]
IBM BLEU (2)

- **Modified** $N$-gram precision between machine translation output and reference translation.
  - Usually with $n$-grams of size 1 to 4

- Modified $n$-gram precision on the entire corpus
  
  $$
P_n = \frac{\sum_{C \in \{Candidates\}, n-gram \in C} \sum \text{count}_{clip} (ngram)}{\sum_{C' \in \{Candidates\}, n-gram' \in C'} \sum \text{count}_{clip} (ngram')}
  $$

- Brevity penalty for too short translations.
  
  $$
  BP = \begin{cases} 
  1 & \text{if } c > r \\
  e^{1 - \frac{r}{c}} & \text{otherwise}
  \end{cases}
  $$
  
  - Typically computed over the entire corpus, not single sentences.
IBM BLEU (3)

- The remote control of the Wii helps to diagnose an infantile ocular disease.
- The control of the Wii helps to diagnose an ocular infantile disease.
- \( w_n = 1/4 \)

\[
BLEU = BP \cdot \exp \left( \sum_{n=1}^{N} w_n \log P_n \right)
\]

<table>
<thead>
<tr>
<th></th>
<th>( p_n )</th>
<th>( BP \cdot p_n )</th>
<th>( w_n \log P_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram precision</td>
<td>13/13 = 1.0</td>
<td>0.926</td>
<td>0</td>
</tr>
<tr>
<td>2-gram precision</td>
<td>8/12 = 0.667</td>
<td>0.617</td>
<td>-0.405</td>
</tr>
<tr>
<td>3-gram precision</td>
<td>6/11 = 0.545</td>
<td>0.505</td>
<td>-0.606</td>
</tr>
<tr>
<td>4-gram precision</td>
<td>5/10 = 0.5</td>
<td>0.463</td>
<td>-0.693</td>
</tr>
<tr>
<td>Brevity penalty</td>
<td>0.926</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( P_n )</td>
<td>Same, only one sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU score</td>
<td>0.6046</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Problems of lexical similarities (1)

- The reliability of lexical metrics depends strongly on the heterogeneity/representativity of reference translations.

- Actually, human translations tend to score low on BLEU.

- Underlying Cause:
  - Lexical similarity is neither a sufficient nor a necessary condition so that two sentences convey the same meaning.
Problems of lexical similarities (2)

- Statistical MT systems heavily rely on the training data.
- Testsets tend to be similar (domain, register, sublanguage) to training materials.
- $N$-gram based metrics favour MT systems which closely replicate the lexical realization of the references.
- Statistical MT systems tend to share the reference sublanguage and be favoured by $n$-gram-based measures.
Overview

- Introduction
- Automatic MT evaluation
- Linguistically motivated evaluation measures
- Quality estimation
- The Asiya toolkit
Linguistically motivated measures (1)

- Extending Lexical Similarity Measures to increase robustness [Gim09]
  - Lexical variants:
    - Morphological information (i.e., stemming) ROUGE and METEOR
    - Synonymy lookup: METEOR (based on WordNet)
  - Paraphrasing support:
    - Extended versions of METEOR, TER
  - Equivalent reference translation graph:
    - HyTER [Dre12]
METEOR (1) [Bano05]

- Parameterized harmonic mean of word P and R
  \[ F_{\text{mean}} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R} \]
- Matching algorithm
  - Exact matching
  - Partial credit for matching stems
  - Partial credit for matching synonyms

- N-gram penalty based on the number of chunks with longer length of adjacent words matched in both strings
  \[ Pen = \gamma \cdot \left( \frac{ch}{m} \right)^\beta \]

- Final score:
  \[ \text{METEOR} = (1 - Pen) \cdot F_{\text{mean}} \]
Extensions

- **METEOR-NEXT**
  - Weighted matches depending on the type
  - Phrase-level matches
  - new matching algorithm accounting for start-positions distance

**Paraphrasing**

- Paraphrase tables from parallel corpora
- Used by the paraphrase matcher

- $\delta$ parameter: content vs. function words discrimination
More linguistically-motivated measures

- Features capturing **syntactic** and **semantic** information
  - Shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, discourse representation, error categories, ...

- Some linguistically-motivated measures:
  - IQmt [Gim09] – syntactic and semantics
  - MaxSim [Chao8] - syntactic
  - RTE [Pado9] – textual entailment
  - VERTa [Com14] – syntactic and semantics
Example 1: Structural Similarity (1)

- Rather than comparing sentences at lexical level: *Compare the linguistic structures* and the words within them [Gim10]

- Compare different linguistic-level elements
  - Words, lemmas, POS, Chunks
  - Parsing Trees
  - Named entities and semantic roles
  - Discourse representation (logical forms)
Example 1: Structural Similarity (2)

- The remote control of the Wii helps to diagnose an infantile ocular disease.

- The Wii remote helps diagnose a childhood eye disease.
Example 1: Structural Similarity (3)
Example 1: Structural Similarity (4)
Example 1: Structural Similarity (4)
Measuring structural similarity (1)

- Linguistic Element (LE): abstract reference to any possible type of linguistic unit, structure, or relationship among them.
  - For instance: POS tags, word lemmas, NPs, semantic roles, dependency relations, etc.

- A sentence can be seen as a bag (or a sequence) of LEs of a certain type
Measuring structural similarity (2)

- **OVERLAP** [Gimo07]: generic similarity measure among linguistic elements inspired by the Jaccard coefficient [Jac1901]

- **SEMPOS** [Mac08] is a MT evaluation measure that considers several overlapping variations

- **MATCHING** is a more strict variant [Gim10]
  - All items inside an element are considered the same unit.
  - Computes the proportion of fully translated LEs according to their types.
Overlap (1)

\[ O(t) = \frac{\sum_{i \in (\text{items}_{t}^{\text{cand}} \cap \text{items}_{t}^{\text{ref}})} \text{count}_{\text{cand}}(i,t)}{\sum_{i \in (\text{items}_{t}^{\text{cand}} \cup \text{items}_{t}^{\text{ref}})} \max(\text{count}_{\text{cand}}(i,t), \text{count}_{\text{ref}}(i,t))} \]

\[ O(*) = \frac{\sum_{t \in T} \sum_{i \in (\text{items}_{t}^{\text{cand}} \cap \text{items}_{t}^{\text{ref}})} \text{count}_{\text{cand}}(i,t)}{\sum_{t \in T} \sum_{i \in (\text{items}_{t}^{\text{cand}} \cup \text{items}_{t}^{\text{ref}})} \max(\text{count}_{\text{cand}}(i,t), \text{count}_{\text{ref}}(i,t))} \]
The remote control of the Wii helps to diagnose an infantile ocular disease.

The Wii remote helps diagnose a childhood eye disease.

Overlap:
- Intersection: 13
- Union: 25
- Ol = 13/25 = 0.52
The remote control of the Wii helps to diagnose an infantile ocular disease.

The Wii remote helps diagnose a childhood eye disease.

Overlap:
- Intersection: 9
- Union: 15
- $\text{Ol} = \frac{9}{15} = 0.6$

<table>
<thead>
<tr>
<th>Words</th>
<th>Reference</th>
<th>Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>JJ</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>NN</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>IN</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NNP</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>VBZ</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TO</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>VB</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
CL-Explicit Semantic Analysis

- CL-ESA requires a significant comparable corpus $C_I$

- $d_q \in L \ (d' \in L')$ is represented as a vector of relations to the index collection $C_I \ (C_I')$

- Monolingual similarities are computed over the VSM (e.g., the cosine of the vocabulary) [Poto8]
Example 2: Semantic Analysis (2)
Towards Heterogeneous MT Evaluation

Lexical Recall
Lexical Precision
F-measure
PoS Tagging
Dependency Parsing
Named Entities
Semantics Roles
Chunking
Constituency Parsing
Discourse Representations

Lexical Similarity
Syntactic Similarity
Semantic Similarity
Towards Heterogeneous MT Evaluation

**Lexical Similarity**
- WNM
- Lexical Precision
- ROUGE
- SIA
- METEOR
- BLEU
- CDER
- GTM
- BLANC
- PER
- WER

**Syntactic Similarity**
- SP-Op-*
- PoS Tagging
- DP-Op-*
- DP-Or-*
- SP-NISTp
- BERT
- MAXSIM
- SP-NISTc
- SP-NISTI
- CP-Op-*
- CP-Oc-*
- STM

**Semantic Similarity**
- DP-OI-*
- Dependency Parsing
- DP-Oc-*
- HWCM
- NE-Oe-*
- NEE
- NE-Me-*
- SR-Or-*
- SR-Mr-*
- DR-Or-*
- Discourse Representations

**Lexical Entities**
- NER

MT Tutorial

46
Metric Combination

- Different measures capture different aspects of similarity

- Simple Approach
  - ULC: Uniformly-averaged linear combination of measures

- But, which ones?
  - Simple hill climbing approach to find the best subset of measures $M$ on a development corpus
    - $M = \{\text{ROUGE}_w, \text{METEOR}, \text{DP-HWC}_r, \text{DP-O}_c(\ast), \text{DP-O}_l(\ast), \text{DP-O}_r(\ast), \text{CP-STM}_r, \text{SR-O}_c(\ast), \text{SR-O}_s, \text{DR-O}_p(\ast) \}$
The goal is to **combine the scores** conferred by different evaluation measures **into a single measure of quality** such that their relative contribution is adjusted on the basis of human feedback (i.e. from human assessments).

**Examples:**
- AMBER [Che12] – downhill simplex
- SIMBLEU (ROSE) [Son11] – SVM
- SPEDE [Wan12] – pFSM for regression
- TERRORCAT [Fis12] - SVM on error categories
Overview

- Introduction
- Automatic MT evaluation
- Linguistically motivated evaluation measures
- **Quality estimation**
- The Asiya toolkit
Quality Estimation (1)

- Setting:
  - Quality assessment without reference translations

- Information available:
  - Source sentence, candidate translation(s) and, possibly, MT system information

- Motivation:
  - System ranking (system selection)
  - Hypotheses re-ranking (parameter optimization)
  - Feedback filtering (especially end-users)
  - Post-edition effort (industry pricing)
Quality Estimation (2)

- Relevant Work:

- Recent work:
  - (Specia et al., 2009;2010), (Soricut and Echihabi, 2010), (Giménez and Specia 2010), (Pighin et al., 2011), (Avramidis, 2012)

- WMT shared task on Quality Estimation
  - [Cal12] WMT12 – 11 participants
  - [Boj13] WMT13 – 14 participants
  - (3d edition at WMT 2014)
Quality Estimation Features (1)

- **System-dependent**
  - internal system probabilities/scores (automatic score)
  - features over *n-best translation hypotheses*
    - language modelling
    - candidates rank
    - score ratio
    - average candidates length
    - length ratio
    - ...

---

MT Tutorial
Quality Estimation Features (2)

- **System-independent**
  - source (translation difficulty)
    - Source sentence length
    - Ambiguity $\rightarrow$ dictionary/alignment/WordNet-based
      - e.g., number of candidate translations per word or phrase
  
  - target (fluency)
    - OOV
    - Language models: perplexity, log probability
Quality Estimation Features (3)

- **System-independent**
  - source-target (adequacy)
    - length factor
    - punctuation and symbols concurrency
    - candidate matching \(\rightarrow\) dictionary-/alignment-based
    - character \(n\)-grams [McNo4]
    - pseudo-cognates [Sim92]
    - word alignments [Gon14]
QE Challenges

- QE is a difficult task
  - Few corpus available
  - Too domain-oriented

<table>
<thead>
<tr>
<th>DE-EN, task 1.2, QE2013</th>
<th>Kendall’s τ ties ignored</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFKI-logregFss33</td>
<td>0.31</td>
</tr>
<tr>
<td>DFKI-logregFss24</td>
<td>0.28</td>
</tr>
<tr>
<td>UPC-1</td>
<td>0.27</td>
</tr>
<tr>
<td>UPC-2</td>
<td>0.24</td>
</tr>
<tr>
<td>DCU-CCG</td>
<td>0.18</td>
</tr>
<tr>
<td>CNGL-SVRPLSF1</td>
<td>0.17</td>
</tr>
<tr>
<td>CNGL-SVRF1</td>
<td>0.17</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.08</td>
</tr>
<tr>
<td>Oracle BLEU</td>
<td>0.22</td>
</tr>
<tr>
<td>Oracle METEOR-ex</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Overview

- Introduction
- Automatic MT evaluation
- Linguistically motivated evaluation measures
- Quality estimation
- The Asiya toolkit
Asiya is an Open Toolkit for Automatic Machine Translation and (Meta-)Evaluation

http://asiya.lsi.upc.edu

Asiya provides:
- Automatic evaluation measures using several linguistic layers for a variety of languages
- Quality Estimation measures
- Meta-evaluation metrics
- Learning schemes
Asiya

- **Languages:**
  - English, Spanish, Catalan
  - Czech, French, German and Russian with limited resources

- **Similarity principles**
  - Precision, Recall, Overlap, Matching, ...

- **Linguistic layers:**
  - Lexical, Syntactic, Semantic, Discourse
Metrics and Meta-metrics

813 metrics are available for language ‘es’ -> ‘en’
Asiya how to (1)

- Asiya operates over testbeds (or test suites).
  - a testbed is a collection of test cases:
    - Source segment
    - Candidate translation(s)
    - Reference translation(s)
Asiya how to (2)

- Asiya.pl Asiya.config
- Asiya.config:

```plaintext
# MANDATORY DEFINITIONS
input=raw
srclang=es
srccase=cs
trglang=en
trgcase=cs

#SOURCE definition ================
src=./data/src.txt

#REFERENCE definition =============
ref=./data/ref.txt

#SYSTEM OUTPUT definition =========
sys=./data/google.txt
sys=./data/bing.txt
sys=./data/reverso.txt
sys=./data/systran.txt
sys=./data/aperium.txt
sys=./data/habelfish.txt
```

MT Tutorial
Asiya how to (3)

- General Options
  - Input format
    - Raw
    - Nist
  - Language pair
    - Srclang
    - Trglang
  - Predefined sets of metrics, systems and references

MT Tutorial
Asiya Interfaces
Hands-on

http://asiya.lsi.upc.edu

- Choose the languages
- Write some sentences or upload a SMALL file. Try to introduce several errors:
  - lexical disagreement, missing prepositions,
- Use some linguistic measures in addition to the lexical ones:
  - BLEU, NIST, ROUGE\textsubscript{w}, METEOR-pa
  - SP-Op(*), DP-HWC, DP-Or(*), CP-STM
- Run it and look how the segment level scores identify the errors in each sentence
- Look at the parse trees
- Use the tSearch interface to find interesting sentences according to the scores and the parse trees
References
[LDC05] NIST Multimodal Information Group. NIST 2005 Open Machine Translation (OpenMT)


References


Additional Slides
Evaluation of syntactic measures

- NIST 2005 Arabic-to-English Exercise

<table>
<thead>
<tr>
<th>Level</th>
<th>Metric</th>
<th>$\rho_{all}$</th>
<th>$\rho_{SMT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>BLEU</td>
<td>0.06</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>0.05</td>
<td>0.90</td>
</tr>
<tr>
<td>Syntactic</td>
<td>POS</td>
<td>0.42</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>DP</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.74</td>
<td>0.95</td>
</tr>
<tr>
<td>Semantic</td>
<td>SR</td>
<td>0.72</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>DR-POS</td>
<td>0.97</td>
<td>0.90</td>
</tr>
</tbody>
</table>