Labelled Dependencies in Machine Translation Evaluation

Karolina Owczarzak, Josef van Genabith, Andy Way
{owczarzak,josef,away}@computing.dcu.ie
National Centre for Language Technology, School of Computing, Dublin City University

Automatic MT metrics: fast and cheap way to evaluate MT systems

- string-based techniques compare the surface form of the translation sentence to the surface form of the reference sentence(s)
- criticised for their insensitivity to perfectly legitimate syntactic and lexical variation between the translation and the reference
- almost all attempts at creating better metrics limited to accomodating a degree of paraphrasing and/or surface reordering of elements, while still ignoring the structural level

Example

John resigned yesterday vs. *Yesterday, John quit*

1-grams: 2/3 (john, resigned, yesterday)

2-grams: 0/2 (john resigned, resigned yesterday)

3-grams: 0/1 (john resigned yesterday)

2/6 n-grams total

BLEU score (Papineni et al., 2002): 33% correct Human score: 100% correct

Automatic MT metrics: variations on string-based comparison

BLEU (Papineni et al., 2002):

number of shared n-grams, brevity penalty

NIST (Doddington, 2002):

number of shared n-grams weighted by frequency, brevity penalty

General Text Matcher (GTM) (Turian et al., 2003):

precision and recall on translation-reference pairs, weights contiguous matches more than non-contiguous matches

Translation Error Rate (TER) (Snover et al., 2006):

edit distance for translation-reference pair, number of insertions, deletions, substitutions and shifts; human-assisted version HTER requires editing of references

METEOR (Banerjee and Lavie, 2005):

sum of n-gram matches for exact string forms, stemmed words, and WordNet synonyms

Kauchak and Barzilay (2006): using WordNet synonyms with BLEU

Owczarzak et al. (2006): using paraphrases derived from the test set through word/phrase alignment with BLEU and NIST

Dependencies in MT Evaluation

Liu and Gildea (2005):

calculating number of matches on syntactic features and unlabelled dependencies; dependencies are non-labelled head-modifier sequences derived by head-extraction rules from syntactic trees.

This work:

follows and extends Liu and Gildea (2005); precision and recall on labelled dependencies extracted with an LFG parser.

Labelled Dependencies

Predicate dependencies: adjunct, apposition, complement, open complement, coordination, determiner, object, second object, oblique, second oblique, oblique agent, possessive, quantifier, relative clause, subject, topic, relative clause pronoun

Non-predicate dependencies: adjectival degree, coordination surface form, focus, if, whether, that, modal, number, verbal particle, participle, passive, person, pronoun surface form, tense, infinitival clause

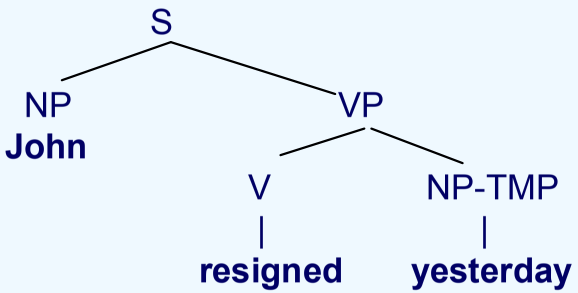
Lexical-Functional Grammar (LFG)

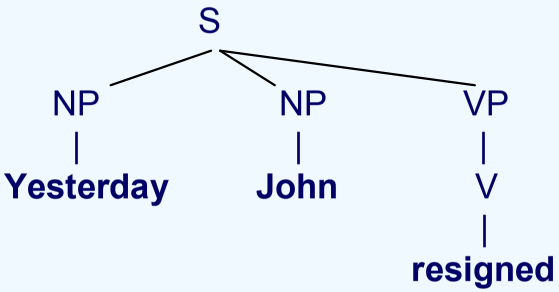
Sentence structure representation in LFG:

- c-structure (constituent): CFG trees, reflects surface word order and structural hierarchy
- f-structure (functional): abstract grammatical (syntactic) relations

John resigned yesterday vs. *Yesterday, John resigned*

c-structure level:





f-structure level:

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{john} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix} \\ \text{PRED} & \text{resign} \\ \text{TENSE} & \text{past} \\ \text{ADJ} & \{ [\text{PRED} \quad \text{yesterday}] \} \end{bmatrix}$$

vs.

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{john} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix} \\ \text{PRED} & \text{resign} \\ \text{TENSE} & \text{past} \\ \text{ADJ} & \{ [\text{PRED} \quad \text{yesterday}] \} \end{bmatrix}$$

= 100% MATCH

The LFG Parser

- Cahill et al. (2004): LFG parser based on Penn II Treebank
- demo at http://lfg-demo.computing.dcu.ie/lfgparser.html

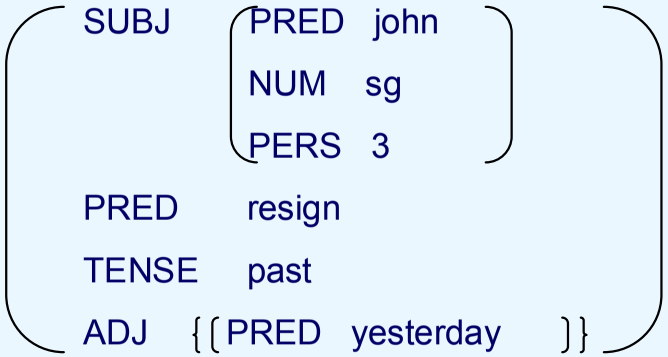
Evaluation of parser quality as MT evaluation

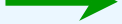
- parser quality is evaluated by comparing the dependencies produced by the parser with the set of dependencies in human annotation of same text in terms of precision, recall, and f-score

- the same can be used to evaluate the quality of translation: Parse the translation and the reference into LFG f-structures rendered as dependency triples, calculate precision, recall, and f-score for the translation-reference pair

Dependencies

Labelled dependency triples are a flat format in which f-structures can be presented.





triples:

SUBJ(resign, john)
PERS(john, 3)
NUM(john, sg)
TENSE(resign, past)
ADJ(resign, yesterday)
PERS(yesterday, 3)
NUM(yesterday, sg)

triples – predicates only:

SUBJ(resign, john)
ADJ(resign, yesterday)

Determining the level of parser noise

- 100 English sentences hand-modified: changing the placement of the adjunct, the order of coordinated elements, no change in meaning or grammaticality
- change limited to c-structure, no change in f-structure
- a perfect parser should give both identical set of dependencies, i.e. the f-score should be perfect

Example:

Schengen, on the other hand, is not organic. original "reference"
On the other hand, Schengen is not organic. modified "translation"

Result:

To alleviate parser noise, we can use a number of best parses on each side of the comparison (translation and reference) – this should eliminate most accidental parsing mistakes.

number of parses	dependencies f-score	predicates-only f-score
perfect parser	100	100
50 best	98.79	97.63
30 best	98.74	X
20 best	98.59	X
10 best	98.31	X
5 best	97.90	X
2 best	97.31	X
1 best	96.56	94.13

Correlation with human judgement - experiment

- 5,007 segments randomly selected from LDC Chinese-English Multiple Translation (parts 2&4)
- each segment: translation + reference + human scores for fluency and accuracy
- evaluated with BLEU, NIST, GTM, METEOR, TER, a number of versions of labelled dependency-based method

Versions of labelled dependency-based method:

- n-best parses on each side of the comparison (translation and reference) to alleviate parser noise (1, 2, 10, 50 best)

- addition of WordNet to compare with WordNet-enhanced version of METEOR

- all dependencies, predicate-only dependencies (ignoring "atomic" features such as *person*, *number*, *tense*, etc., or a subset excluding irrelevant dependency types (as determined in training) such as *if*, *modal*, *obl*, *to_inf*, *topic*, *whether*, *xcomp*, *tense*, *that*

- partial matching for predicate dependencies, to score cases, where one correct lexical object happens to find itself in the correct relation, but with an incorrect "partner"

subj (resign , John) → subj (resign , x) , subj (y , John)

- weights for each dependency type (as determined in training) to maximize the correlation of combined score to human judgements; combined with default proportional weights

- change proportion of precision and recall in total f-score

Correlation with human judgement – results

fluency	
d_50_var_WN_excl	0.184832256
d_50_var_WN	0.181609767
d_50	0.165599242
d_fl_prop	0.163996627
d_var	0.161794771
d_WN	0.159882714
d_excl	0.154419408
METEOR+WN	0.153589228
d	0.152946914
d_excl_pr	0.148589839
METEOR	0.145159211
GTM	0.143461381
TER	0.142005976
BLEU	0.141443164
d_fluency	0.140627531
NIST	0.139595414

accuracy	
METEOR+WN	0.291327377
d_50_var_WN_excl	0.291052646
d_50_var_WN	0.285762786
METEOR	0.272374525
NIST	0.268528171
d_var	0.267712505
d_WN	0.261838877
GTM	0.259921382
d_50	0.258908101
d_excl_pr	0.257655907
d_excl	0.256832799
d	0.253963183
d_acc_prop	0.223642773
d_accuracy	0.206723231
BLEU	0.194602661
TER	0.192966354

average	
d_50_var_WN_excl	0.267044037
d_50_var_WN	0.262255506
METEOR+WN	0.252376531
d_var	0.241727866
d_50	0.238111562
d_WN	0.23720207
METEOR	0.236699473
NIST	0.231688212
d_excl	0.23152421
d_excl_pr	0.229262842
d	0.229057442
GTM	0.228222829
d_av_prop	0.212400648
d_average	0.191529622
BLEU	0.187032065
TER	0.186291468

d: dependency method baseline version; **d_var:** partial matching; **d_50:** 50 best parses; **d_WN:** WordNet; **d_fluency, d_accuracy, d_average:** added weights for individual dependency types (as determined in training) to maximize correlation with fluency, accuracy, or average human score; **d_fl_prop, d_acc_prop, d_av_prop:** as above, but with proportional weights as well (depending on frequency of given dependency type within segment); **d_excl:** excluding irrelevant dependency types; **d_excl_pr:** as above, but with f-score calculated from weighted precision (0.3) and recall (1.7); **d_50_var_WN_excl:** 50 best parses, partial matching, WordNet; **d_50_var_WN_excl:** 50 best parses, partial matching, WordNet, excluded 0-weight dependency types

Correlation with human judgement – discussion

- correlation with human fluency judgements much lower for all metrics than with accuracy judgements

- our method outperforms others at reflecting fluency judgements, comparable to METEOR at reflecting accuracy judgements

  - the dependency-based method is very sensitive to the grammatical structure of the sentence: a more grammatical translation is also a translation that is more fluent

  - METEOR or NIST assign relatively little importance to the position of a specific word in a sentence, therefore they are more sensitive to content rather than linguistic form

- fluency and accuracy – two very different aspects of translation quality, each with its own set of conditions along which the input is evaluated; a single automatic metric unlikely to correlate highly with human judgements of both at the same time (see GTM and METEOR)

- adding the partial matching option in our method = greatest increase in correlation (the partial-match versions consistently outperformed versions with a larger number of parses available but without the partial match)

- the partial-match versions (even with a single parse) offered results comparable to or higher than the addition of WordNet to the matching process for accuracy and overall judgement.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization: 65-73.

Joan Bresnan. 2001. *Lexical-Functional Syntax*, Blackwell, Oxford.

Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef van Genabith, and Andy Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. Proceedings of ACL 2004: 320-327.

George Doddington. 2002. Automatic Evaluation of MT Quality using N-gram Co-occurrence Statistics. Proceedings of HLT 2002: 138-145.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.

Kaplan, Ronald M. and Joan Bresnan. 1982. *Lexical-functional Grammar: A Formal System for Grammatical Representation*. In J. Bresnan (ed.), The Mental Representation of Grammatical Relations. MIT Press, Cambridge.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. Proceedings of HLT-NAACL 2006: 45-462.

Karolina Owczarzak, Declan Groves, Josef van Genabith, and Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. Proceedings of the HLT-NAACL 2006 Workshop on Statistical Machine Translation: 86-93.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of ACL 2002: 311-318.

Mathew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciula. 2006. A Study of Translation Error Rate with Targeted Human Annotation. Proceedings of AMTA 2006: 223-231.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and Its Evaluation. *Proceedings of MT Summit 2003*: 386-393. New Orleans, Louisiana.

DCU
Dublin City University
Ollscoil Chathair Bhaile Átha Cliath

Microsoft

ENTERPRISE IRELAND