Dependency-Based Automatic Evaluation for Machine Translation

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 your MT system

The quality of Machine Translation (MT) output is usually evaluated by string-based techniques, which compare the surface form of the translation sentence to the surface form of the reference sentence(s).

The coT o f t sig-□ c

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; their 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

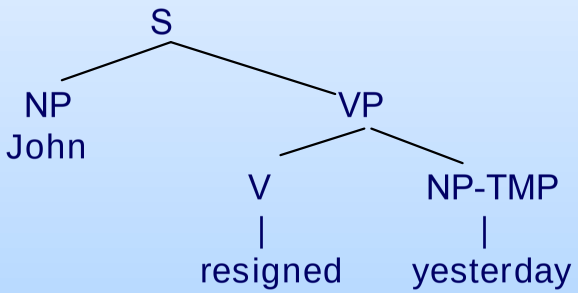
Lexical-Functional Grammar (LFG)

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:

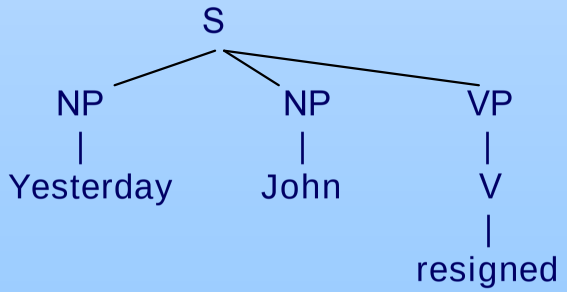


f-structure level:

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

vs.

= 100% MATCH



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

The LFG Parser

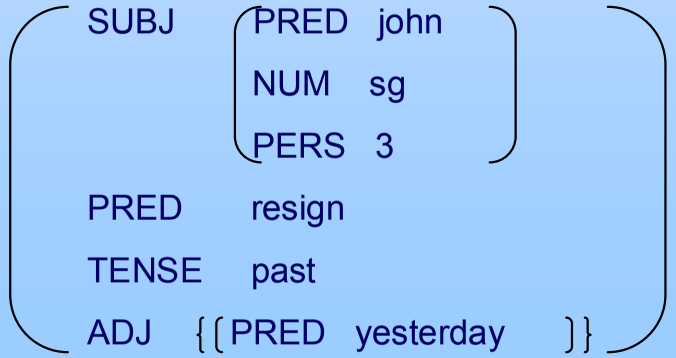
Cahill et al. (2004) presents an LFG parser based on Penn II Treebank (demo at http://lfg-demo.computing.dcu.ie/lfgparser.html). It automatically annotates Charniak's or Bikel's output parse with attribute-value equations and resolves to f-structures. High precision and recall, provides a parse in 99.9% of cases.

Evaluation of parser quality as MT evaluation

The quality of the parser can be determined by comparing the dependencies produced by the parser with the set of dependencies in human annotation of same text, and calculating precision, recall, and f-score. The same process 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 to change the placement of the adjunct or 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

16,807 segments from LDC Chinese-English Multiple Translation project, parts 2 and 4. Each segment consists of translation, reference, and 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 or predicate-only dependencies (ignoring “atomic” features such as *person*, *number*, *tense*, etc.

- 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)

Correlation with human judgement – results

fluency		accuracy		average	
d_50+WN	0.177	M+WN	0.294	M+WN	0.255
d+WN	0.175	M	0.278	d_50_var	0.252
d_50_var	0.174	d_50_var	0.273	d_50+WN	0.25
GTM	0.172	NIST	0.273	d_10_var	0.25
d_10_var	0.172	d_10_var	0.273	d_2_var	0.247
d_50	0.171	d_2_var	0.27	d+WN	0.244
d_2_var	0.168	d_50+WN	0.269	d_50	0.243
d_10	0.168	d_var	0.266	d_var	0.243
d_var	0.165	d_50	0.262	M	0.242
d_2	0.164	d_10	0.262	d_10	0.242
d	0.161	d+WN	0.26	NIST	0.238
BLEU	0.155	d_2	0.257	d_2	0.237
M+WN	0.153	d	0.256	d	0.235
M	0.149	d_pr	0.24	d_pr	0.216
NIST	0.146	GTM	0.203	GTM	0.208
d_pr	0.143	BLEU	0.199	BLEU	0.197
TER	0.133	TER	0.192	TER	0.182

d = dependency f-score, _pr = predicate-only f-score, 2, 10, 50 = n-best parses; var = partial-match version; M = METEOR, WN = WordNet

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, but is not the best 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 those with just 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, Luisiana.

Microsoft

ENTERPRISE IRELAND