Alignment-based sense selection in METEOR and the RATATOUILLE recipe

Benjamin Marie[1,2] *& Marianna Apidianaki*[1]
(1) LIMSI-CNRS, (2) Lingua et Machina

METEOR-WSD

Why WSD ?

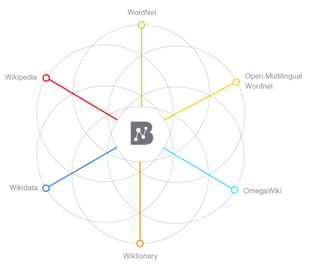
- Meteor and Meteor-NEXT map words with identical surface forms or having the same stem, WordNet synonyms and paraphrases but all variants are treated as semantically equivalent ⇒ synonyms and paraphrases of different senses mapped during evaluation
- Meteor-WSD: identify the correct set of synonyms/paraphrases for a word/phrase in context

Alignment-based MultiWord Expression (MWE) identification

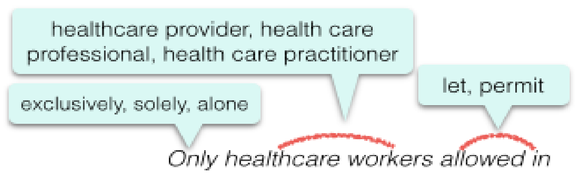
- WMT'14 data lemmatised, PoS tagged and word aligned
- candidate MWE: a sequence of words in the reference that is aligned to a single source word (n:1) (e.g. *téléphone portable* - *cellphone*)
- MWE validated if it exists in BabelNet; its variants are retrieved (e.g. *téléphone cellulaire*, *GSM*, *téléphone mobile*)

Alignment-based Disambiguation

- identify the meaning of words in the reference sentences using alignments and the multilingual semantic network BabelNet (Apidianaki and Gong, 2015; Navigli and Ponzetto, 2012)



- find BabelNet synsets describing the senses of a word w and containing its aligned translation in a precise context
- keep the variants provided for w in the retained synsets
- fall back to the most frequent sense (synset) for unaligned English words or when the translation is not found in any synset
- reference sentences enriched with variants valid in this context



Results

- at segment-level (metric: Kendall's τ, data: WMT'14)

Metric	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
Meteor-1.5	.406	.334	.282	.329	.338	.280	.238	.318	.427	.316	.327
Meteor-WSD	.410	.332	.282	.332	.339	.280	.240	.321	.437	.320	.330

- at system-level (metric: Pearson's coefficient, data: WMT'14)

Metric	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
Meteor-1.5	.975	.927	.980	.805	.922	.941	.263	.976	.923	.776	.849
Meteor-WSD	.975	.927	.979	.828	.927	.946	.258	.981	.929	.779	.852

RATATOUILLE

A metric combination

- combine Meteor-WSD and nine other metrics: PER, WER, CDER, TER, GTM 1.3, sentence-level BLEU, Meteor 1.5, RIBES 1.03.1 and BEER 1.0

- each metric gives a score at segment-level
- RATATOUILLE is the result of the log-linear combination of each metric's score

Tuning

- the weight for each metric score is tuned using a similar approach to PRO (Hopkins and May, 2011; Guzmán et al., 2014)
- a pairwise approach: candidate translation pairs are classified into 2 categories (correctly or incorrectly ordered)
- training examples: all translation pairs or only translation pairs of significant different quality separated by at least 3 ranks in the human judgments
- at segment-level, experiments show a slightly better correlation when using only translation pairs of significant different quality

Tuning set	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
`all`	.426	.336	.294	.337	.348	.292	.286	.352	.459	.347	.348
`>=3`	.425	.342	.297	.340	.351	.293	.292	.345	.456	.347	.349

Results and importance of Meteor-WSD

- at segment-level (metric: Kendall's τ, data: WMT'14)

Metric	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
BEER	.417	.337	.284	.333	.343	.292	.268	.344	.440	.336	.340
RATATOUILLE w/o Meteor-WSD	.423	.343	.296	.338	.350	.293	.291	.344	.454	.346	.348
RATATOUILLE w/o Meteor-1.5	.425	.341	.297	.339	.351	.293	.292	.345	.458	.347	.349
RATATOUILLE	.425	.342	.297	.340	.351	.293	.292	.345	.456	.347	.349

- at system-level, RATATOUILLE score for each segment is first passed through a sigmoid function and the final system score is the average of all segment scores (metric: Pearson's coefficient, data: WMT'14)

Metric	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
Meteor 1.5	.975	.927	.980	.805	.922	.941	.263	.976	.923	.776	.849
RATATOUILLE w/o Meteor-WSD	.974	.900	.994	.804	.918	.955	.403	.979	.946	.821	.869
RATATOUILLE w/o Meteor-1.5	.974	.899	.993	.804	.918	.958	.408	.979	.945	.823	.870
RATATOUILLE	.974	.901	.993	.804	.918	.959	.408	.979	.944	.823	.870

- at segment-level RATATOUILLE outperforms BEER, the best metric from WMT'14
- at both system-level and segment-level RATATOUILLE gives slightly better results with Meteor-WSD than with Meteor-1.5

CONCLUSION AND FUTURE WORK

- context-dependent sense selection helps Meteor establish better correspondences between hypotheses and references and improves the performance of the RATATOUILLE metric in almost all language pairs
- in the future, we intend to perform context-based filtering of pivot paraphrases and replace Meteor by Meteor-WSD in BEER to improve its correlation with human judgments

WMT'15 OFFICIAL RESULTS AT SEGMENT-LEVEL

Direction / Extracted-pairs	fr-en 29770	fi-en 31577	de-en 40535	cs-en 85877	ru-en 44539	Average
DPMFCOMB	.395 ± .012	**.445** ± .012	**.482** ± .009	**.495** ± .007	**.418** ± .013	**.447** ± .011
BEER_TREEPEL	.389 ± .014	.438 ± .010	.447 ± .008	.471 ± .007	.403 ± .014	.429 ± .011
RATATOUILLE	**.398** ± .010	.421 ± .011	.441 ± .010	.472 ± .007	.393 ± .013	.425 ± .010
UPF-COBALT	.386 ± .012	.437 ± .013	.427 ± .011	.457 ± .007	.402 ± .013	.422 ± .011
BEER	.393 ± .012	.422 ± .012	.438 ± .010	.457 ± .008	.396 ± .014	.421 ± .011
CHRF	.383 ± .011	.417 ± .012	.424 ± .010	.446 ± .008	.384 ± .014	.411 ± .011
CHRF3	.383 ± .013	.397 ± .011	.421 ± .010	.449 ± .008	.386 ± .013	.407 ± .011
METEOR-WSD	.375 ± .012	.406 ± .010	.420 ± .011	.438 ± .008	.387 ± .012	.405 ± .010
DPMF	.368 ± .012	.411 ± .011	.418 ± .011	.436 ± .008	.378 ± .011	.402 ± .011
LEBLEU-OPTIMIZED	.376 ± .013	.391 ± .010	.399 ± .010	.438 ± .008	.374 ± .012	.396 ± .011
LEBLEU-DEFAULT	.373 ± .013	.383 ± .011	.402 ± .009	.436 ± .007	.376 ± .011	.394 ± .010
VERTA-EQ	.388 ± .012	.369 ± .013	.410 ± .011	.447 ± .007	.346 ± .013	.392 ± .011
VERTA-70ADEQ30FLU	.374 ± .012	.365 ± .014	.418 ± .011	.438 ± .007	.344 ± .013	.388 ± .011
VERTA-W	.383 ± .010	.344 ± .014	.416 ± .010	.445 ± .007	.345 ± .013	.387 ± .011
DREEM	.362 ± .012	.340 ± .010	.368 ± .011	.423 ± .007	.348 ± .013	.368 ± .011
UOW-LSTM	.332 ± .011	.376 ± .012	.375 ± .011	.385 ± .008	.356 ± .010	.365 ± .011
SENTBLEU	.358 ± .013	.308 ± .012	.360 ± .011	.391 ± .006	.329 ± .011	.349 ± .011
TOTAL-BS	.332 ± .013	.319 ± .013	.333 ± .010	.381 ± .007	.321 ± .011	.337 ± .011
USAAR-ZWICKEL-METEOR	n/a	.406 ± .011	.422 ± .011	.439 ± .008	.386 ± .012	.413 ± .011
USAAR-ZWICKEL-COMET	n/a	.021 ± .013	.050 ± .010	.072 ± .009	.084 ± .010	.057 ± .011
USAAR-ZWICKEL-COSINE2METEOR	n/a	.001 ± .013	−.011 ± .010	.020 ± .009	.041 ± .010	.013 ± .011
USAAR-ZWICKEL-COSINE	n/a	−.035 ± .013	−.019 ± .010	.090 ± .008	.014 ± .013	.012 ± .011

Direction / Extracted-pairs	en-fr 34512	en-fi 32694	en-de 54447	en-cs 136890	en-ru 49302	Average
BEER	.352 ± .010	**.380** ± .010	.393 ± .010	.435 ± .006	**.439** ± .010	**.400** ± .009
CHRF3	.335 ± .013	.373 ± .012	.398 ± .008	**.446** ± .005	.420 ± .010	.395 ± .010
RATATOUILLE	**.366** ± .013	.318 ± .011	.381 ± .008	.429 ± .006	.436 ± .010	.386 ± .010
LEBLEU-OPTIMIZED	.347 ± .009	.368 ± .010	**.399** ± .008	.410 ± .006	.404 ± .011	.386 ± .009
CHRF	.342 ± .012	.359 ± .010	.372 ± .010	.444 ± .005	.410 ± .011	.385 ± .010
LEBLEU-DEFAULT	.345 ± .010	.368 ± .010	.398 ± .009	.406 ± .006	.404 ± .012	.384 ± .009
METEOR-WSD	.342 ± .012	.286 ± .010	.344 ± .007	.390 ± .006	.399 ± .010	.352 ± .009
DREEM	.338 ± .012	.280 ± .011	.317 ± .010	.395 ± .006	.366 ± .010	.339 ± .010
SENTBLEU	.318 ± .011	.227 ± .011	.294 ± .009	.360 ± .005	.347 ± .010	.309 ± .009
TOTAL-BS	.297 ± .011	.223 ± .009	.278 ± .009	.345 ± .005	.356 ± .011	.300 ± .009
DPMF	.335 ± .012	n/a	.350 ± .009	n/a	n/a	.343 ± .010
USAAR-ZWICKEL-METEOR	n/a	n/a	.342 ± .008	n/a	n/a	.342 ± .008
USAAR-ZWICKEL-COMET	n/a	n/a	.056 ± .019	n/a	n/a	.056 ± .009
USAAR-ZWICKEL-COSINE	n/a	n/a	−.007 ± .010	n/a	n/a	−.007 ± .010
USAAR-ZWICKEL-COSINE2METEOR	n/a	n/a	−.027 ± .019	n/a	n/a	−.027 ± .009