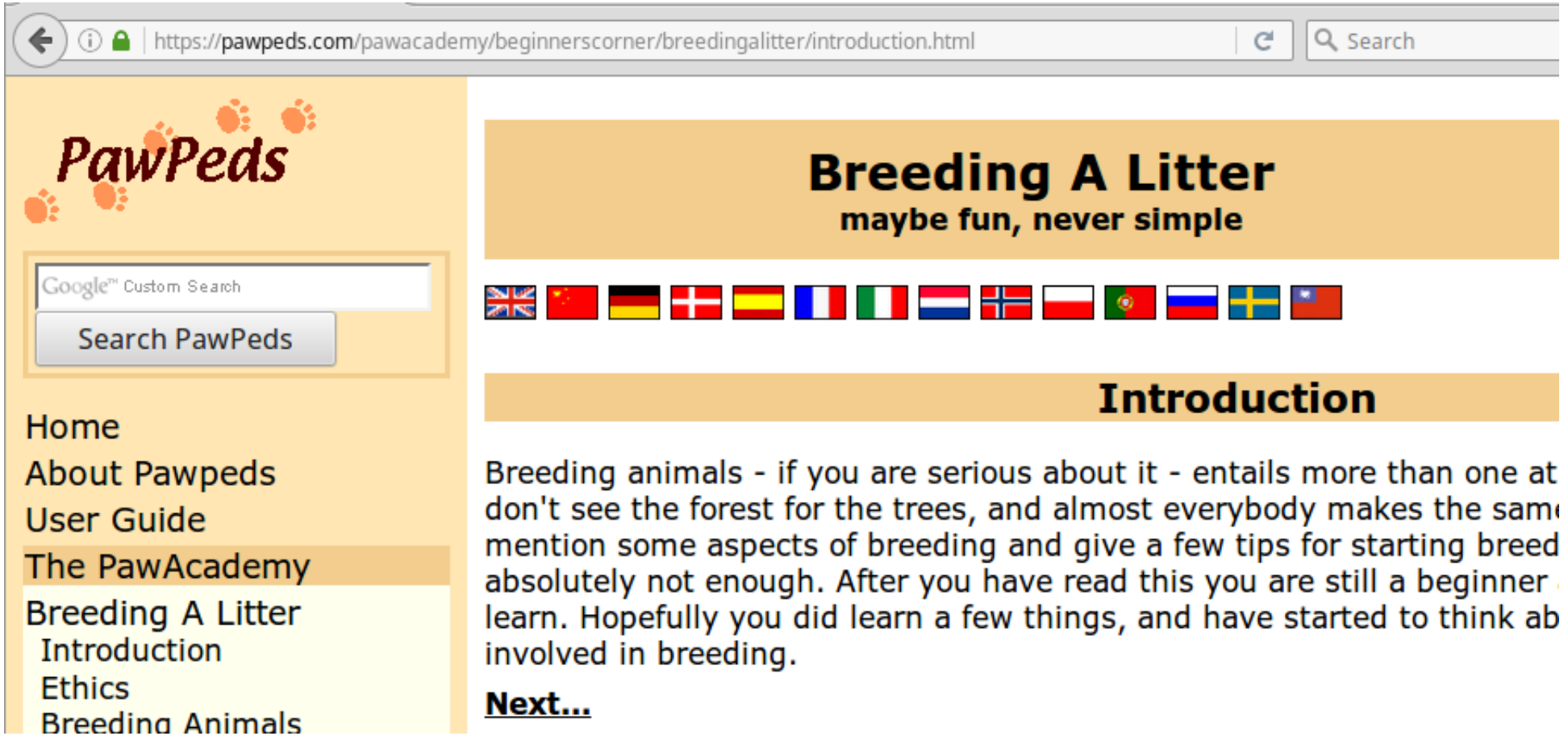
# Shared Task
# Bilingual Document Alignment

Christian Buck and Philipp Koehn
University of Edinburgh / Johns Hopkins University

12 August 2016


# Document Alignment

Finding pairs of documents that are translations of each other



# Document Alignment

Finding pairs of documents that are translations of each other

# Document Alignment

Finding pairs of documents that are translations of each other

# Motivation

MT training data

- There's no data like more data
- BLEU goes up
- Different effects on big / small data

Previous work

- Scattered efforts
- No common evaluation

# Data

Training

- 1,624 English-French pairs
- From 49 webdomains
- Between 4 and 200 per webdomain

Test

- 2402 English-French pairs
- From 203 new webdomains

# Preparation steps provided to participants



- Download HTML files (using HTTrack)
- Fix encoding issues
- Detection of document language (using CLD2)
- Text extraction (using HTML5 parser)
- Translation of French text to English (using, of course, Moses)
- Easy file format (thanks, Bitextor) + Python examples
- Baseline: `green.com/fr_FR/witch-fr == green.com/witch`

# Evaluation & 1-1 Rule

- Recall only
- BUT: 1-1 rule; **every document can only occur in one pair**
- URL-matching baseline: 60% recall

# Challenges

Big-ish websites

- E.g. `cinedoc.org`: 50k English, 50k French pages
- Makes 2.5B possible pairs
- Only allowed to pick 50k

Language detection unreliable

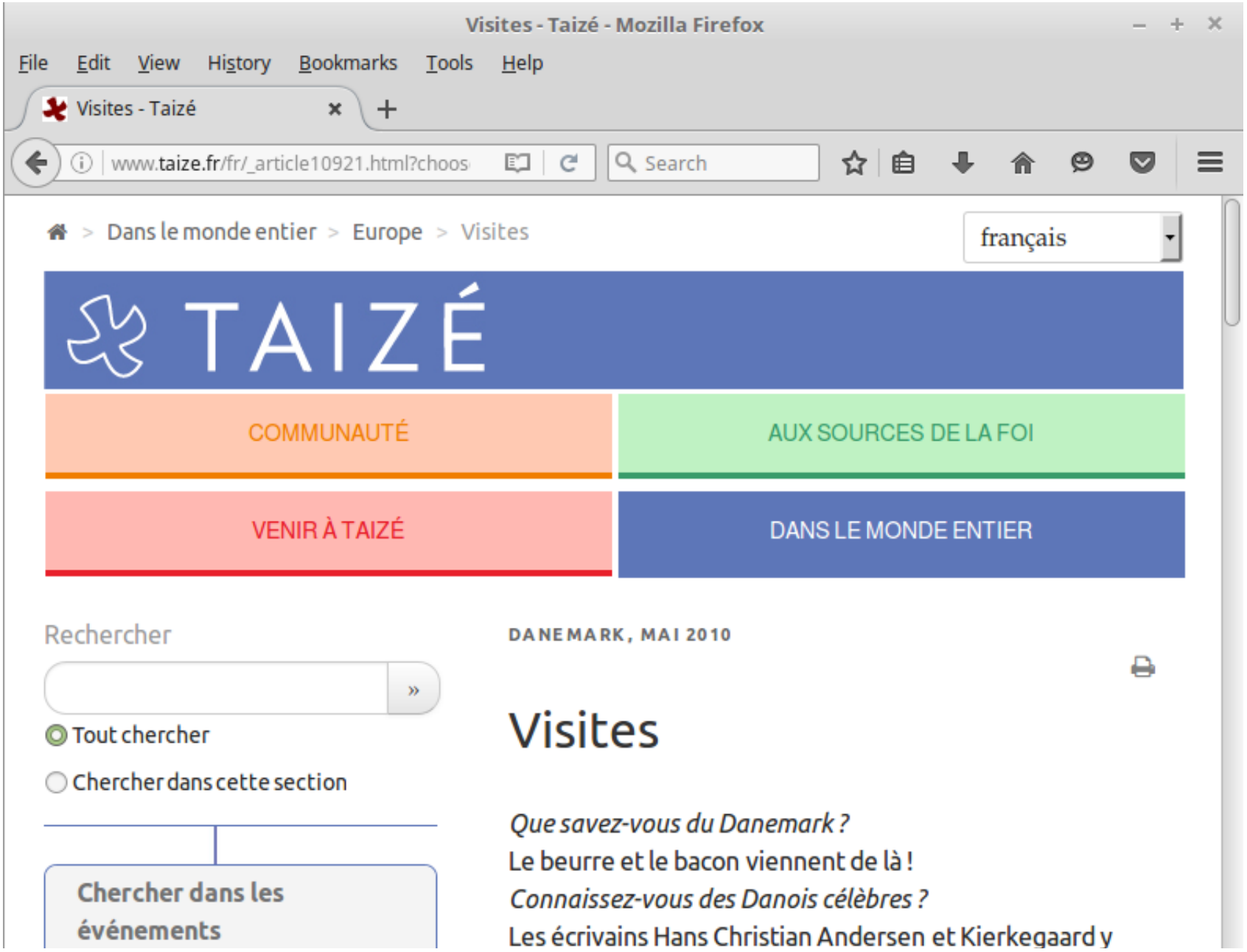
- Made sure test set can be found
- Some participants ran their own pipelines

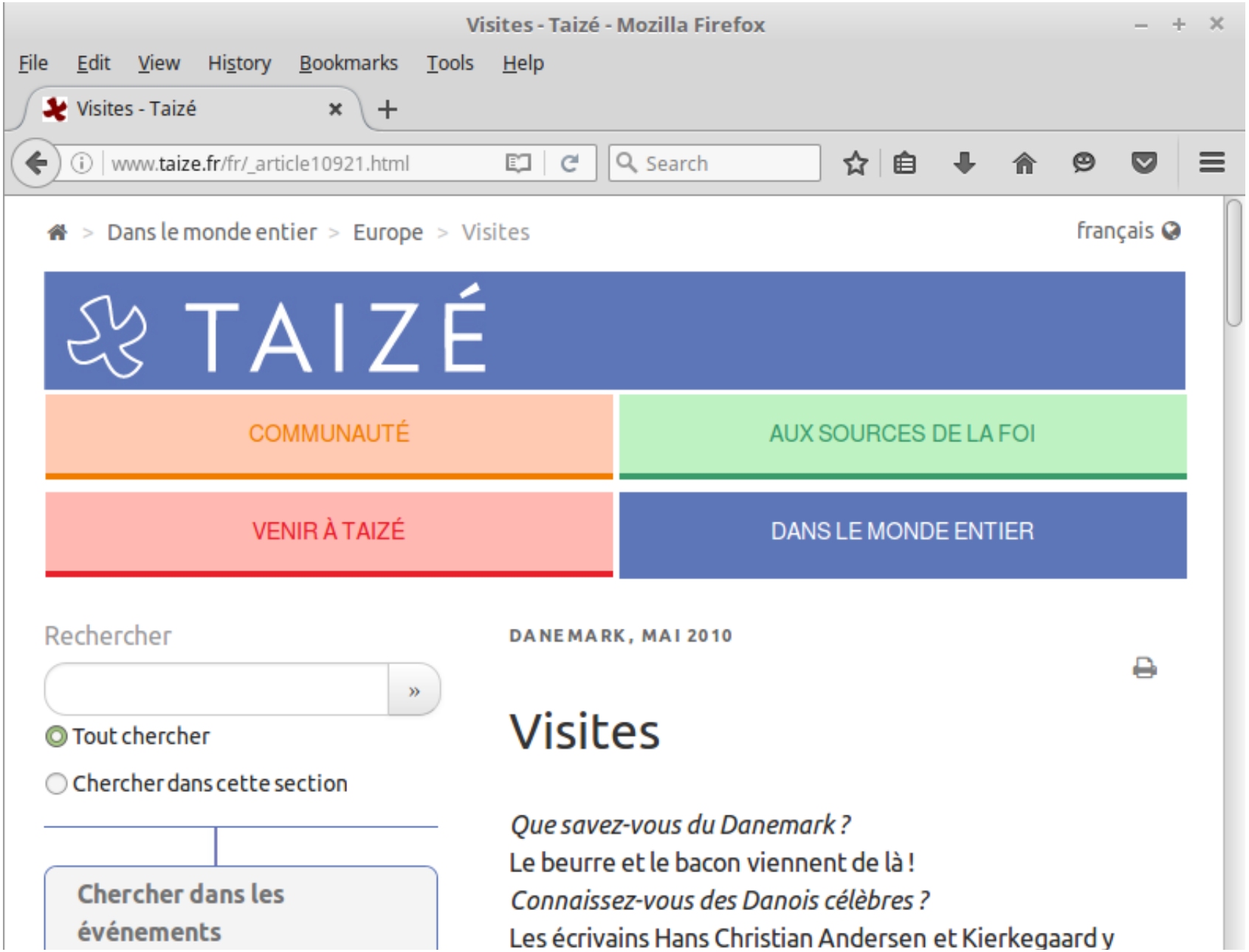
# Challenges II



Near duplicates

- Removed pages when text was *exactly* the same
- `www.taize.fr/fr_article10921.html`
- `www.taize.fr/fr_article10921.html?chooselang=1`
- *Almost* identical


Visites - Taizé - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Visites - Taizé
www.taize.fr/fr/_article10921.html?choos
Search
> Dans le monde entier > Europe > Visites
français
TAIZÉ
COMMUNAUTÉ
AUX SOURCES DE LA FOI
VENIR À TAIZÉ
DANS LE MONDE ENTIER
Rechercher
»
Tout chercher
Chercher dans cette section
Chercher dans les événements
DANEMARK, MAI 2010
Visites
Que savez-vous du Danemark ?
Le beurre et le bacon viennent de là !
Connaissez-vous des Danois célèbres ?
Les écrivains Hans Christian Andersen et Kierkegaard y


Visites - Taizé - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Visites - Taizé
www.taize.fr/fr/_article10921.html
Search
Dans le monde entier > Europe > Visites
français
TAIZÉ
COMMUNAUTÉ
AUX SOURCES DE LA FOI
VENIR À TAIZÉ
DANS LE MONDE ENTIER
Rechercher
Tout chercher
Chercher dans cette section
Chercher dans les événements
DANEMARK, MAI 2010
Visites
Que savez-vous du Danemark ?
Le beurre et le bacon viennent de là !
Connaissez-vous des Danois célèbres ?
Les écrivains Hans Christian Andersen et Kierkegaard y

# Results!



- 11 participating groups
- 19 submissions
- Up to 95% recall (NovaLincs-URL-Coverage)

| Name | Predicted pairs | Pairs after 1-1 rule | Found pairs | Recall % |
|---|---|---|---|---|
| ADAPT | 61 094 | 61 094 | 644 | 26.8 |
| ADAPT-v2 | 69 518 | 69 518 | 651 | 27.1 |
| BadLuc | 681 610 | 263 133 | 1 905 | 79.3 |
| DOCAL | 191 993 | 191 993 | 2 128 | 88.6 |
| ILSP-ARC-pv42 | 291 749 | 287 860 | 2 040 | 84.9 |
| JIS | 323 929 | 28 903 | 48 | 2.0 |
| Medved | 155 891 | 155 891 | 1 907 | 79.4 |
| NovaLincs-coverage-url | 207 022 | 207 022 | 2 060 | 85.8 |
| NovaLincs-coverage | 235 763 | 235 763 | 2 129 | 88.6 |
| **NovaLincs-url-coverage** | 235 812 | 235 812 | 2 281 | 95.0 |
| UA PROMPSIT bitextor 4.1 | 95 760 | 95 760 | 748 | 31.1 |
| UA PROMPSIT bitextor 5.0 | 157 682 | 157 682 | 2 001 | 83.3 |
| UEdin1 cosine | 368 260 | 368 260 | 2 140 | 89.1 |
| UEdin2 LSI | 681 744 | 271 626 | 2 062 | 85.8 |
| UEdin2 LSI-v2 | 367 948 | 367 948 | 2 105 | 87.6 |
| UFAL-1 | 592 337 | 248 344 | 1 953 | 81.3 |
| UFAL-2 | 574 433 | 178 038 | 1 901 | 79.1 |
| UFAL-3 | 574 434 | 207 358 | 1 938 | 80.7 |
| UFAL-4 | 1 080 962 | 268 105 | 2 023 | 84.2 |
| YSDA | 277 896 | 277 896 | 2 021 | 84.1 |
| YODA | 318 568 | 318 568 | 2 256 | 93.9 |
| Baseline | 148 537 | 148 537 | 1 436 | 59.8 |

# Allowing 5% edits between predicted and expected



| Name | Pairs found | Δ | Recall | Δ | Rank | Δ |
|---|---|---|---|---|---|---|
| ADAPT | 726 | +82 | 30.2 | +3.4 | 20 | 0 |
| ADAPT-v2 | 733 | +82 | 30.5 | +3.4 | 19 | 0 |
| BadLuc | 2 062 | +157 | 85.9 | +6.5 | 13 | +3 |
| DOCAL | 2 235 | +107 | 93.1 | +4.5 | 4 | +1 |
| ILSP-ARC-pv42 | 2 185 | +145 | 91.0 | +6.0 | 7 | +2 |
| JIS | 48 | 0 | 2.0 | 0.0 | 21 | 0 |
| Medved | 1 986 | +79 | 82.7 | +3.3 | 15 | 0 |
| NovaLincs-coverage-url | 2 130 | +70 | 88.7 | +2.9 | 9 | −1 |
| NovaLincs-coverage | 2 192 | +63 | 91.3 | +2.6 | 6 | −2 |
| NovaLincs-url-coverage | 2 303 | +22 | 95.9 | +0.9 | 2 | −1 |
| UA PROMPSIT bitextor 4.1 | 775 | +27 | 32.3 | +1.1 | 18 | 0 |
| UA PROMPSIT bitextor 5.0 | 2 117 | +116 | 88.1 | +4.8 | 10 | +2 |
| UEdin1 cosine | 2 227 | +87 | 92.7 | +3.6 | 5 | −2 |
| UEdin2 LSI | 2 146 | +84 | 89.3 | +3.5 | 8 | −1 |
| UEdin2 LSI-v2 | 2 281 | +176 | 95.0 | +7.3 | 3 | +3 |
| UFAL-1 | 2 060 | +107 | 85.8 | +4.5 | 14 | −1 |
| UFAL-2 | 1 954 | +53 | 81.4 | +2.2 | 17 | 0 |
| UFAL-3 | 1 980 | +42 | 82.4 | +1.8 | 16 | −2 |
| UFAL-4 | 2 078 | +55 | 86.5 | +2.3 | 12 | −2 |
| YSDA | 2 102 | +81 | 87.5 | +3.4 | 11 | 0 |
| **YODA** | 2 307 | +51 | 96.0 | +2.1 | 1 | +1 |

# Insights



- Machine translated text helpful
- Finding matching n-grams works well
- Big boost by combination with URL-matching baseline
- Content based > structural features?

# thank you