Inferring syntactic rules for word alignment through Inductive Logic Programming

Sylwia Ozdowska, Vincent Claveau

CLLE-ERSS - Univ. of Toulouse
Toulouse, France

IRISA-CNRS
Rennes, France

May 19, 2010
Word alignment

Definition and use

- link occurrences of words (or phrases) that are in a translation relationship in parallel corpora
- usefulness of word alignment (Véronis 00)
  - acquisition of bilingual lexical resources, machine translation, cross-lingual information retrieval...

Existing techniques

- most approaches:
  - statistical alignment models (Brown et al. 93)
  - lexicon-based alignment models (Gale & Church 91)
- growing interest for syntax-informed models (Wu 00; Yamada & Knight 01; Gildea 03; Lin & Cherry 03)
Syntax and alignment

Debili & Zribi’s hypothesis (96)
- if two words are translations of each other in aligned sentences, then their respective governors and dependents may be translations of each other

ALIBI (Ozdowska, 06)
- rule-based system for English/French
- principle: from two aligned anchor words (AW), the alignment link is projected to syntactically connected words
Introduction

Syntax and alignment

Debili & Zribi's hypothesis (96)

- if two words are translations of each other in aligned sentences, then their respective governors and dependents may be translations of each other

ALIBI (Ozdowska, 06)

- rule-based system for English/French
- principle: from two aligned anchor words (AW), the alignment link is projected to syntactically connected words

\[
\text{The } \underline{\text{Community}} \quad \text{banned} \quad \text{imports of ivory} \\
\quad I \\
\text{La } \underline{\text{Communauté}} \text{ a interdit l’importation d’ivoire}
\]
Syntax and alignment

Debili & Zribi’s hypothesis (96)

- if two words are translations of each other in aligned sentences, then their respective governors and dependents may be translations of each other

ALIBI (Ozdowska, 06)

- rule-based system for English/French
- principle: from two aligned anchor words (AW), the alignment link is projected to syntactically connected words

\[
\begin{align*}
\text{The } & \text{ Community } \text{ banned } \text{ imports of ivory} \\
\text{La } & \text{ Communauté a interdit l’importation d’ivoire}
\end{align*}
\]
Debili & Zribi’s hypothesis (96)

- if two words are translations of each other in aligned sentences, then their respective governors and dependents may be translations of each other.

**ALIBI** (Ozdowska, 06)

- rule-based system for English/French
- principle: from two aligned anchor words (AW), the alignment link is projected to syntactically connected words.

\[
\begin{array}{c}
\text{The } \underline{\text{Community}} \quad \underline{\text{banned}} \quad \text{imports of ivory} \\
\vdots \\
\text{La } \underline{\text{Communauté}} \quad \underline{\text{a interdit}} \quad \text{l’importation d’ivoire}
\end{array}
\]
Syntax and alignment

Syntactic propagation rules

- key component of the alignment system
- isomorphism (identical syntactic path): V-subj-N / V-subj-N

```
The Community banned imports of ivory
```

```
La Communauté a interdit l’importation d’ivoire
```

Ozdowska, Claveau (ERSS / IRISA)

ILP for alignment
Syntax and alignment

Syntactic propagation rules

- key component of the alignment system
- isomorphism (identical syntactic path): V-subj-N / V-subj-N
- non-isomorphism (compatible pattern): V-obj-N / V-pp+pcomp-N

\[ \ldots \text{affects} \quad \text{cell} \quad \text{stability} \]

\[ \ldots \text{intervient sur la stabilité des cellules} \]
Syntax and alignment

Syntactic propagation rules

- key component of the alignment system
- isomorphism (identical syntactic path): V-subj-N / V-subj-N
- non-isomorphism (compatible pattern): V-obj-N / V-pp+pcomp-N

Manual-encoding of the rules

- yields good results...
- ... yet defining these propagation rules is an issue
  - necessitate an expert in both languages
  - tedious task to be carried out for any new pair of languages, of parsers...

⇒ machine learning of the propagation rules
Supervised approach

- examples are pairs of words, linked by a syntactic path in both languages

Inductive Logic Programming (ILP)

- highly expressive, symbolic ML technique (Muggleton 95)
  - examples and output in first order logic (Prolog)
- natural way to encode relations and external knowledge
  - eg. translation and syntactic relations with simple predicates:
    \[ x \text{ is the subject of } y = \text{subj}(x, y) \]
- outputs human readable rules, making a linguistic analysis possible
Machine learning of alignment rules

Inductive Logic Programming

Theoretical framework of ILP

- infer a set of rules $H$ (Horn clauses)...
- ...from examples $E^+$ (and possibly counter-examples $E^-$)
- ...and a Background Knowledge $B$
- ...such as $B \land H \land E^- \not\models \Box$ and $B \land H \models E^+$

In our case

- $H$: syntactic propagation rules
- $E^+$: pairs of AW (no counter-examples)
- $B$: dependency relations and AW
In practice

- training data
  - aligned sentence
    
    \[
    \text{private sector companies} / \ les \ entreprises \ du \ secteur \ privé
    \]
    
    \[
    e_1 \quad e_2 \quad e_3 \quad f_1 \quad f_2 \quad f_3 \quad f_4 \quad f_5
    \]
  
  - dependency relations and AW in \( B \)
    
    \[
    \begin{align*}
    \text{adj}(e_2,e_1). & \quad \text{det}(f_2,f_1). & \quad \text{pcomp}(f_3,f_4). & \quad \text{aw}(e_2,f_4). \\
    \text{nn}(e_3,e_2). & \quad \text{pp}(f_2,f_3). & \quad \text{adj}(f_4,f_5). & \quad \text{aw}(e_3,f_2). 
    \end{align*}
    \]

- several rules generated for each example, organized in a lattice
  - for ex., align(E,F) :- \( \text{nn}(E,E_2), \text{pp}(F,F_3), \text{pcomp}(F_3,F_4), \text{aw}(E_2,F_4). \)

\[
\begin{array}{cccc}
\ldots & \boxed{E_2} & \ldots & E & \ldots \\
\ldots & F & \ldots & F_3 & \ldots & \boxed{F_4} & \ldots \\
\end{array}
\]

\[
\begin{array}{lll}
\text{nn} & \text{pp} & \text{pcomp}
\end{array}
\]
Machine learning of alignment rules

Search lattice built on one example
- each rule of the lattice is scored wrt the other examples
- the best one is kept in $H$

align(E,F).

align(E,F) :- nn(E,E1).

align(E,F) :- pcomp(F3,F4), aw(E2,F4).

align(E,F) :- nn(E,E2), pp(F,F3),

align(E,F) :- nn(E,E2), pp(F,F3),
pcomp(F3,F4), aw(E2,F4).

align(E,F) :- adj(E2,E1), det(F,F1), aw(E2,F4), nn(E,E2), pp(F2,F3), aw(E,F), adj(F4,F5), pcomp(F,F4) ...
The whole picture

Alignment algorithm

1. generate the examples: anchoring
   - cognates: string similarity (Fluhr et al. 00)
   - lexicon: simple cooccurrence model (Gale & Church 92)

2. parse the bitext
   - Syntex FR and Syntex EN (Bourigault 07)

3. infer propagation rules with ILP
   - ALEPH implementation (Srinivasan 01)

4. apply the rules to any bitext (after parsing and anchoring)

5. consider found alignments as anchors and goto 4
Experiments

Questions about ILP
- performance for the alignment task?
- interpretability of the inferred rules?

Questions about training data
- influence of the type of the training corpus?
- influence of the size of the training corpus?
Performance evaluation

Evaluation framework

- **training dataset**
  - HANSARD corpus (RALI, Univ. of Montreal)
  - Canadian parliamentary debates
  - 1000 sentences used for the training

- **test set: HLT’03 dataset**
  - 447 sentences from the Hansards (≠ training corpus)
  - sure alignments S (inter-annotator agreement on S) and probable alignments P (multi-word expressions, free translations...)

- evaluation in precision (P), recall (R) and f-measure (F)
Performance evaluation

Results on S alignments from HLT’03 data set

<table>
<thead>
<tr>
<th>System</th>
<th>ALIBI</th>
<th>ILP</th>
<th>Ralign</th>
<th>XRCE</th>
<th>BiBr</th>
<th>ProAlign</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.89</td>
<td>0.82</td>
<td>0.72</td>
<td>0.55</td>
<td>0.63</td>
<td>0.72</td>
</tr>
<tr>
<td>R</td>
<td>0.67</td>
<td>0.74</td>
<td>0.81</td>
<td>0.93</td>
<td>0.74</td>
<td>0.91</td>
</tr>
<tr>
<td>F</td>
<td>0.76</td>
<td>0.78</td>
<td>0.76</td>
<td>0.69</td>
<td>0.68</td>
<td>0.80</td>
</tr>
</tbody>
</table>

- Performance comparable with existing alignment systems (Mihalcea & Pedersen 03)
  - higher P
  - lower R
Performance evaluation

Cause of errors

Misalignments

- mostly caused by parsing errors
  - adjective *federal* was wrongly attached to *carpenters* leading to the misalignment *carpenter* / *gouvernement* in *federal government carpenters get* $6.42 / *Les menuisiers du gouvernement fédéral touchent* $6.42.

- caused by overgeneralization
  - *gouvernement* and *legislation* are misaligned in the sentence pair: *good legislation has been brought in by Liberal governments* / *les gouvernements libéraux ont apporté de bonnes mesures législatives*.

Non detected alignments

- lack of anchor pairs and of dependency relations
Corpora

- **HANSARD**
- **INRA**
  - Institut National de la Recherche Agronomique
  - research and popular science articles on agronomy
  - \( \sim 300,000 \) word tokens
- **JOC**
  - ARCADE Project (Véronis & Langlais 00)
  - various questions and answers dealt with at the European Commission
  - \( \sim 400,000 \) word tokens
- 1,000 sentences for each corpus used for training (separately)
Performance on HLT’03 test set

<table>
<thead>
<tr>
<th>Training corpus</th>
<th>HANSARD</th>
<th>JOC</th>
<th>INRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>82.08%</td>
<td>80.65%</td>
<td>83.16%</td>
</tr>
<tr>
<td>R</td>
<td>74.09%</td>
<td>74.10%</td>
<td>66.90%</td>
</tr>
<tr>
<td>F</td>
<td>77.88%</td>
<td>77.20%</td>
<td>74.15%</td>
</tr>
</tbody>
</table>

- Little differences with respect to the type of training corpus (except R on INRA)
- F-measure slightly improves if training and test are done on the same type of corpus
**Inferred rules**

**Genericity**
- \(\sim 60\) rules learned from each corpus of 1000 sentences
  - 38 rules shared across the three corpora
  - 13 to 21 corpus-specific rules

**Comparison with human-generated rules**
- all identical rules encoded in ALIBI were inferred
- most of compatible rules encoded in ALIBI were inferred
- new rules not encoded in ALIBI were found
Size of the training corpus

- 300 to 1000 sentences: little variation in P and R
- < 300: P increases and R decreases
- 10 sentences: 70% f-measure
Concluding remarks

About our approach

- fully automatic approach
  - supervised ML approach bootstrapped by the generation of anchors
- yields good performance
- inferred rules give insights on case of isomorphisms and non-isomorphisms between the two languages

No free lunch

- approach chiefly based on syntax
  - makes the most of knowledge embedded in parsers, thus requires few training data
  - dependent on the existence and quality of the parsers
Perspectives

Improvements

- enrich Background knowledge
  - add information like PoS, lemmas...
  - use the score from statistical alignment approaches
- find strategies to deal with partial syntactic analysis
- extension to dependency tree alignment

Application

- portability to different parsers
- portability to different language pairs
Inferring syntactic rules for word alignment through Inductive Logic Programming

Sylwia Ozdowska, Vincent Claveau

CLLE-ERSS - Univ. of Toulouse
Toulouse, France

IRISA-CNRS
Rennes, France

May 19, 2010
**Parsers**

- **SYNTEX fr** and **SYNTEX en** (Bourigault 07)
  - input: POS tagged sentences (TreeTagger (Schmidt 94))
  - output: dependency relations for each sentence

---

The composition of the medium affects subsequent cell stability

---

La composition du milieu intervient sur la stabilité ultérieure des cellules

---

Both parsers designed according to the same architecture

- Performance: **SYNTEX fr** > **SYNTEX en**
Corpora

- **INRA**
  - Institut National de la Recherche Agronomique
  - research and popular science articles on agronomy
  - $\sim 300 \, 000$ word tokens

- **JOC**
  - ARCADE Project (Véronis & Langlais 00)
  - various questions and answers dealt with at the European Commission
  - $\sim 400 \, 000$ word tokens

- **HANSARD**
  - RALI (University of Montreal)
  - Canadian parliamentary debates
  - $\sim 250 \, 000$ word tokens
Evaluation of overall performance

Method

- Evaluation of precision (P), recall (R) and f-measure (F)
- Cross-corpus evaluation
- Human annotation task
  - 120 test sentences for each corpus
  - annotation guidelines (Melamed 98, Véronis 98)
  - 3 human judges
- Human annotation output for each corpus
  - 60 sentences annotated by 2 persons
  - 60 sentences annotated by 1 person
### Human annotation task

**Inter-annotator agreement estimation**

<table>
<thead>
<tr>
<th></th>
<th>J1J2</th>
<th>J1J3</th>
<th>J2J3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INRA</strong></td>
<td>0.90</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>JOC</strong></td>
<td>0.87</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td><strong>HANSARD</strong></td>
<td>0.76</td>
<td>0.82</td>
<td>0.72</td>
</tr>
</tbody>
</table>

- Overall agreement between pairs of annotators
- Lower agreement on HANSARD than on INRA and JOC
Human annotation task
Different annotation schemes

- Segmentation level

J1  *The [allis shad]_1 [is considered to be]_2 a vulnerable species*
    *La [grande alose]_1 [est considérée comme]_2 une espèce vulnérable*

J2  *The allis\_1 shad\_2 [is considered]\_3 [to be]\_4 a vulnerable species*
    *La grande\_1 alose\_2 [est considérée]\_3 comme\_4 une espèce vulnérable*
Human annotation task
Different annotation schemes

Segmentation level

**J1** The *allis shad*$_1$ *is considered to be*$_2$ a vulnerable species
La *grande alose*$_1$ *est considérée comme*$_2$ une espèce vulnérable

**J2** The *allis*$_1$ *shad*$_2$ *is considered*$_3$ *to be*$_4$ a vulnerable species
La *grande*$_1$ *alose*$_2$ *est considérée*$_3$ *comme*$_4$ une espèce vulnérable

NULL alignments

**J1** … that there is any change *in the balance of ways and means*$_1$
… avoir apporté le moindre changement *au niveau de l’ensemble*$_1$

**J2** … that there is any change *in the balance of ways and means*$_0$
… avoir apporté le moindre changement *au niveau de l’ensemble*$_0$
### Human annotation task

#### Types of correspondences

<table>
<thead>
<tr>
<th></th>
<th>1-1</th>
<th>NULL</th>
<th>chunks</th>
</tr>
</thead>
<tbody>
<tr>
<td>INRA</td>
<td>64%</td>
<td>15%</td>
<td>21%</td>
</tr>
<tr>
<td>JOC</td>
<td>51%</td>
<td>22%</td>
<td>27%</td>
</tr>
<tr>
<td>HANSARD</td>
<td>43%</td>
<td>21%</td>
<td>36%</td>
</tr>
</tbody>
</table>

- **1-1 alignments:** INRA > JOC > HANSARD
- **chunk alignments:** INRA < JOC < HANSARD
### Evaluation of overall performance

#### Results

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INRA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.90 (-0.06)</td>
<td>0.95</td>
<td>0.91 (-0.04)</td>
</tr>
<tr>
<td>R</td>
<td>0.45</td>
<td>0.62 (+0.17)</td>
<td>0.66</td>
<td>0.75 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.61</td>
<td>0.73 (+0.12)</td>
<td>0.78</td>
<td>0.82 (+0.04)</td>
</tr>
<tr>
<td><strong>JOC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.87 (-0.09)</td>
<td>0.93</td>
<td>0.87 (-0.06)</td>
</tr>
<tr>
<td>R</td>
<td>0.43</td>
<td>0.57 (+0.14)</td>
<td>0.58</td>
<td>0.67 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.69 (+0.09)</td>
<td>0.71</td>
<td>0.75 (+0.04)</td>
</tr>
<tr>
<td><strong>HANSARD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.95</td>
<td>0.85 (-0.10)</td>
<td>0.89</td>
<td>0.82 (-0.07)</td>
</tr>
<tr>
<td>R</td>
<td>0.28</td>
<td>0.40 (+0.12)</td>
<td>0.43</td>
<td>0.53 (+0.10)</td>
</tr>
<tr>
<td>F</td>
<td>0.43</td>
<td>0.55 (+0.12)</td>
<td>0.58</td>
<td>0.64 (+0.06)</td>
</tr>
</tbody>
</table>
### Evaluation of overall performance

#### Results

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INRA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.90 (−0.06)</td>
<td>0.95</td>
<td>0.91 (−0.04)</td>
</tr>
<tr>
<td>R</td>
<td>0.45</td>
<td>0.62 (+0.17)</td>
<td>0.66</td>
<td>0.75 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.61</td>
<td>0.73 (+0.12)</td>
<td>0.78</td>
<td>0.82 (+0.04)</td>
</tr>
<tr>
<td><strong>JOC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.87 (−0.09)</td>
<td>0.93</td>
<td>0.87 (−0.06)</td>
</tr>
<tr>
<td>R</td>
<td>0.43</td>
<td>0.57 (+0.14)</td>
<td>0.58</td>
<td>0.67 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.69 (+0.09)</td>
<td>0.71</td>
<td>0.75 (+0.04)</td>
</tr>
<tr>
<td><strong>HANSARD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.95</td>
<td>0.85 (−0.10)</td>
<td>0.89</td>
<td>0.82 (−0.07)</td>
</tr>
<tr>
<td>R</td>
<td>0.28</td>
<td>0.40 (+0.12)</td>
<td>0.43</td>
<td>0.53 (+0.10)</td>
</tr>
<tr>
<td>F</td>
<td>0.43</td>
<td>0.55 (+0.12)</td>
<td>0.58</td>
<td>0.64 (+0.06)</td>
</tr>
</tbody>
</table>
Evaluation of overall performance

Results

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INRA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.90 (−0.06)</td>
<td>0.95</td>
<td>0.91 (−0.04)</td>
</tr>
<tr>
<td>R</td>
<td>0.45</td>
<td>0.62 (+0.17)</td>
<td>0.66</td>
<td>0.75 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.61</td>
<td>0.73 (+0.12)</td>
<td>0.78</td>
<td>0.82 (+0.04)</td>
</tr>
<tr>
<td><strong>JOC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.87 (−0.09)</td>
<td>0.93</td>
<td>0.87 (−0.06)</td>
</tr>
<tr>
<td>R</td>
<td>0.43</td>
<td>0.57 (+0.14)</td>
<td>0.58</td>
<td>0.67 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.69 (+0.09)</td>
<td>0.71</td>
<td>0.75 (+0.04)</td>
</tr>
<tr>
<td><strong>HANSARD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.95</td>
<td>0.85 (−0.10)</td>
<td>0.89</td>
<td>0.82 (−0.07)</td>
</tr>
<tr>
<td>R</td>
<td>0.28</td>
<td>0.40 (+0.12)</td>
<td>0.43</td>
<td>0.53 (+0.10)</td>
</tr>
<tr>
<td>F</td>
<td>0.43</td>
<td>0.55 (+0.12)</td>
<td>0.58</td>
<td>0.64 (+0.06)</td>
</tr>
</tbody>
</table>
## Evaluation of overall performance

### Results

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INRA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.90 (−0.06)</td>
<td>0.95</td>
<td>0.91 (−0.04)</td>
</tr>
<tr>
<td>R</td>
<td>0.45</td>
<td>0.62 (+0.17)</td>
<td>0.66</td>
<td>0.75 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.61</td>
<td>0.73 (+0.12)</td>
<td>0.78</td>
<td>0.82 (+0.04)</td>
</tr>
<tr>
<td><strong>JOC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.87 (−0.09)</td>
<td>0.93</td>
<td>0.87 (−0.06)</td>
</tr>
<tr>
<td>R</td>
<td>0.43</td>
<td>0.57 (+0.14)</td>
<td>0.58</td>
<td>0.67 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.69 (+0.09)</td>
<td>0.71</td>
<td>0.75 (+0.04)</td>
</tr>
<tr>
<td><strong>HANSARD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.95</td>
<td>0.85 (−0.10)</td>
<td>0.89</td>
<td>0.82 (−0.07)</td>
</tr>
<tr>
<td>R</td>
<td>0.28</td>
<td>0.40 (+0.12)</td>
<td>0.43</td>
<td>0.53 (+0.10)</td>
</tr>
<tr>
<td>F</td>
<td>0.43</td>
<td>0.55 (+0.12)</td>
<td>0.58</td>
<td>0.64 (+0.06)</td>
</tr>
</tbody>
</table>
## Evaluation of overall performance

### Results

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INRA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.90 (−0.06)</td>
<td>0.95</td>
<td>0.91 (−0.04)</td>
</tr>
<tr>
<td>R</td>
<td>0.45</td>
<td>0.62 (+0.17)</td>
<td>0.66</td>
<td>0.75 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.61</td>
<td>0.73 (+0.12)</td>
<td>0.78</td>
<td>0.82 (+0.04)</td>
</tr>
<tr>
<td></td>
<td>JOC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.87 (−0.09)</td>
<td>0.93</td>
<td>0.87 (−0.06)</td>
</tr>
<tr>
<td>R</td>
<td>0.43</td>
<td>0.57 (+0.14)</td>
<td>0.58</td>
<td>0.67 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.69 (+0.09)</td>
<td>0.71</td>
<td>0.75 (+0.04)</td>
</tr>
<tr>
<td></td>
<td>HANSARD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.95</td>
<td>0.85 (−0.10)</td>
<td>0.89</td>
<td>0.82 (−0.07)</td>
</tr>
<tr>
<td>R</td>
<td>0.28</td>
<td>0.40 (+0.12)</td>
<td>0.43</td>
<td>0.53 (+0.10)</td>
</tr>
<tr>
<td>F</td>
<td>0.43</td>
<td>0.55 (+0.12)</td>
<td>0.58</td>
<td>0.64 (+0.06)</td>
</tr>
</tbody>
</table>
## Evaluation of overall performances

### Results

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INRA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>0.96</td>
<td>0.90 (−0.06)</td>
<td>0.95</td>
<td>0.91 (−0.04)</td>
</tr>
<tr>
<td><strong>R</strong></td>
<td>0.45</td>
<td>0.62 (+0.17)</td>
<td>0.66</td>
<td>0.75 (+0.09)</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>0.61</td>
<td>0.73 (+0.12)</td>
<td>0.78</td>
<td>0.82 (+0.04)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>JOC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>0.96</td>
<td>0.87 (−0.09)</td>
<td>0.93</td>
<td>0.87 (−0.06)</td>
</tr>
<tr>
<td><strong>R</strong></td>
<td>0.43</td>
<td>0.57 (+0.14)</td>
<td>0.58</td>
<td>0.67 (+0.09)</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>0.60</td>
<td>0.69 (+0.09)</td>
<td>0.71</td>
<td>0.75 (+0.04)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HANSARD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>0.95</td>
<td>0.85 (−0.10)</td>
<td>0.89</td>
<td>0.82 (−0.07)</td>
</tr>
<tr>
<td><strong>R</strong></td>
<td>0.28</td>
<td>0.40 (+0.12)</td>
<td>0.43</td>
<td>0.53 (+0.10)</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>0.43</td>
<td>0.55 (+0.12)</td>
<td>0.58</td>
<td>0.64 (+0.06)</td>
</tr>
</tbody>
</table>
Evaluation of overall performances

Results

<table>
<thead>
<tr>
<th></th>
<th>CLA</th>
<th>ALIBI</th>
<th>GIZA++</th>
<th>ALIBI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>INRA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.90 (-0.06)</td>
<td>0.95</td>
<td>0.91 (-0.04)</td>
</tr>
<tr>
<td>R</td>
<td>0.45</td>
<td>0.62 (+0.17)</td>
<td>0.66</td>
<td>0.75 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.61</td>
<td>0.73 (+0.12)</td>
<td>0.78</td>
<td><strong>0.82 (+0.04)</strong></td>
</tr>
<tr>
<td><strong>JOC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>0.87 (-0.09)</td>
<td>0.93</td>
<td>0.87 (-0.06)</td>
</tr>
<tr>
<td>R</td>
<td>0.43</td>
<td>0.57 (+0.14)</td>
<td>0.58</td>
<td>0.67 (+0.09)</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.69 (+0.09)</td>
<td>0.71</td>
<td><strong>0.75 (+0.04)</strong></td>
</tr>
<tr>
<td><strong>HANSARD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>0.95</td>
<td>0.85 (-0.10)</td>
<td>0.89</td>
<td>0.82 (-0.07)</td>
</tr>
<tr>
<td>R</td>
<td>0.28</td>
<td>0.40 (+0.12)</td>
<td>0.43</td>
<td>0.53 (+0.10)</td>
</tr>
<tr>
<td>F</td>
<td>0.43</td>
<td>0.55 (+0.12)</td>
<td>0.58</td>
<td><strong>0.64 (+0.06)</strong></td>
</tr>
</tbody>
</table>
Evaluation of overall performances

To sum up

- F-measure increased over baselines
- Best strategy: ALIBI bootstrapped with GIZA++
- Much higher performances on INRA than on JOC and HANSARD
  - INRA: $F = 0.82$
  - JOC: $F = 0.75$
  - HANSARD: $F = 0.64$
- HANSARD
  - inter-annotator agreement −−
  - chunk alignments ++
  - performances −−