THE UNIVERSITY of EDINBURGH
informatics

Translating Negation: Induction, Search and Model Errors

Federico Fancellu & Bonnie Webber
School of Informatics
University of Edinburgh
f.fancellu@sms.ed.ac.uk, bonnie@inf.ed.ac.uk

THE UNIVERSITY of EDINBURGH
informatics

Why bother? - Examples

Why bother? - Examples

在 同 一 个 急诊 的 值班 中， 我 两 次 没有 发现 病患 得了 盲肠炎 。

During my emergency duty , I have n’t diagnosed a patient with appendicitis twice .

Why bother? - Examples

在 同 一 个 急诊 的 值班 中， 我 两 次 没有 发现 病患 得了 盲肠炎 。

During my emergency duty , I have n't diagnosed a patient with appendicitis twice .

Why bother? - Examples

在 同 一 个 急诊 的 值班 中， 我 两 次 没有 发现 病患 得了 盲肠炎 。

in an emergency duty , I had two patients have been found .

During my emergency duty , I have n't diagnosed a patient with appendicitis twice .

Why bother? - Examples

在 同 一 个 急诊 的 值班 中， 我 两 次 没有 发现 病患 得了 盲肠炎 。

in an emergency duty , I had two patients have been found .

During my emergency duty , I have n't diagnosed a patient with appendicitis twice .

Why bother? - Examples

在 同 一 个 急诊 的 值班 中 ， 我 两 次 没有 发现 病患 得了 盲肠炎 。

in an emergency duty , I had two patients have been found .

During my emergency duty , I have n’t diagnosed a patient with appendicitis twice .

Why bother? - Examples

在 同 一 个 急诊 的 值班 中 ， 我 两 次 没有 发现 病患 得了 盲肠炎 。

?

in an emergency duty , I had two patients have been found .

During my emergency duty , I have n't diagnosed a patient with appendicitis twice .

THE UNIVERSITY of EDINBURGH
informatics

Why bother? - Examples

Why bother? - Examples

如果 他 拿 了 不 属于 他 的 东西 并不 说明 他 就是 个 惯偷

Even if he took things that do not belong to him, that does n't mean he is a thief

Why bother? - Examples

如果 他 拿 了 不 属于 他 的 东西 并不 说明 他 就是 个 惯偷

Even if he took things that do not belong to him, that does n't mean he is a thief

Why bother? - Examples

如果 他 拿 了 不 属于 他 的 东西 并不 说明 他 就是 个 惯偷

if he took to the things he does not mean that he is not a

Even if he took things that do not belong to him, that does n't mean he is a thief

Why bother? - Examples

如果 他 拿 了 不 属于 他 的 东西 并不 说明 他 就是 个 惯偷

if he took to the things he does not mean that he is not a

Even if he took things that do not belong to him, that does n't mean he is a thief

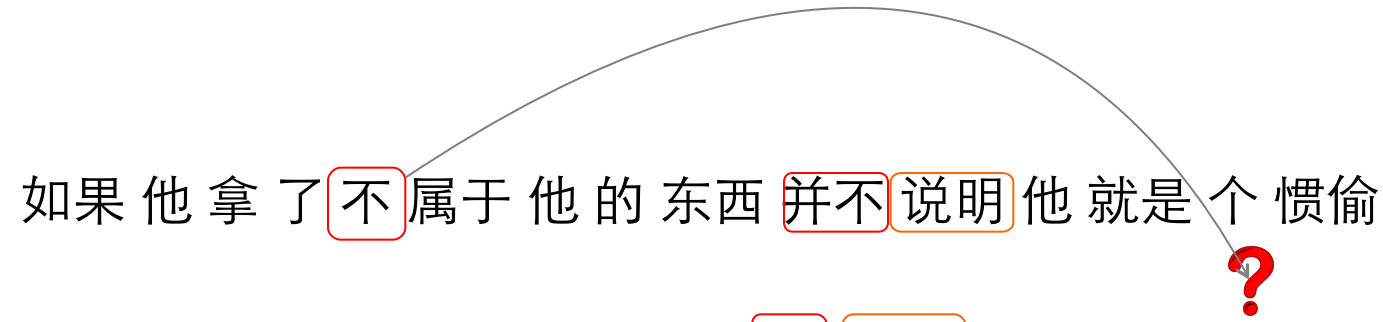
Why bother? - Examples

如果 他 拿 了 不 属于 他 的 东西 并不 说明 他 就是 个 惯偷

if he took to the things he does not mean that he is not a

Even if he took things that do not belong to him, that does n't mean he is a thief

Why bother? - Examples



Even if he took things that do not belong to him, that does n't mean he is a thief

THE UNIVERSITY of EDINBURGH
informatics

Why bother? – BLEU scores

Why bother? – BLEU scores

- BLEU scores also showed a problem in translating negation

Why bother? – BLEU scores

- BLEU scores also showed a problem in translating negation

	Source → Target	Pos	Neg
Wetzel & Bond (2012)	Jp → En	26.70	22.77 (-3.93)
Fancellu & Webber (2014)	Zh → En	27.16	24.3 (-2.86)

Why bother? – BLEU scores

- BLEU scores also showed a problem in translating negation

	Source → Target	Pos	Neg
Wetzel & Bond (2012)	Jp → En	26.70	22.77 (-3.93)
Fancellu & Webber (2014)	Zh → En	27.16	24.3 (-2.86)

- Similar trend for:
  - German → English
  - Czech → English

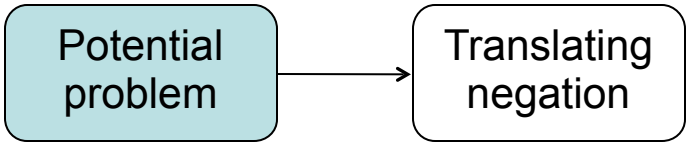
THE UNIVERSITY of EDINBURGH
informatics

What's exactly wrong with negation?

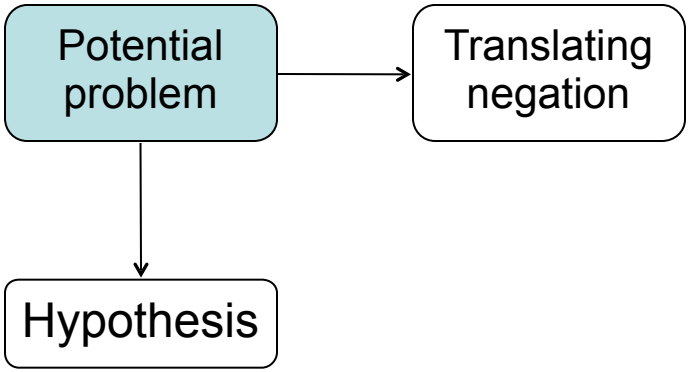
What's exactly wrong with negation?

Potential problem

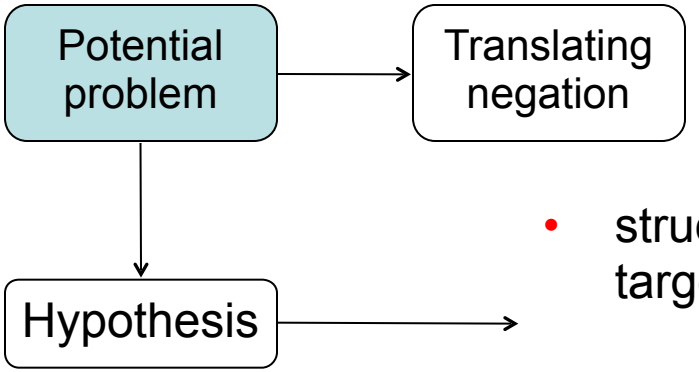
What's exactly wrong with negation?



What's exactly wrong with negation?

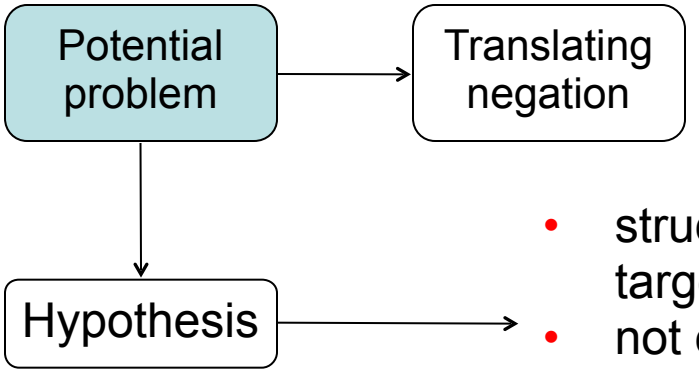


What's exactly wrong with negation?



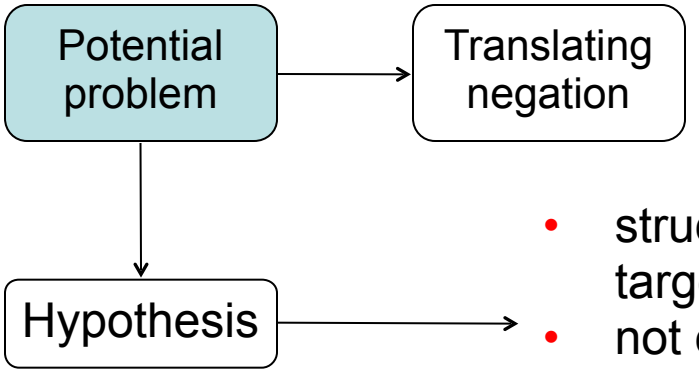
- structural mismatch between source and target language (Collins et al.,2005)

What's exactly wrong with negation?



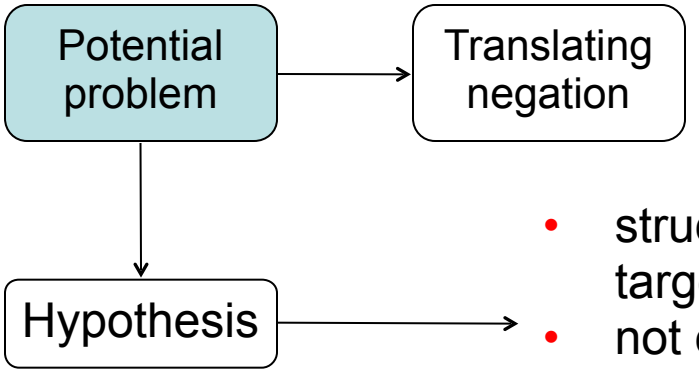
- structural mismatch between source and target language (Collins et al.,2005)
- not enough negative training data (Wetzel & Bond, 2012)

What's exactly wrong with negation?



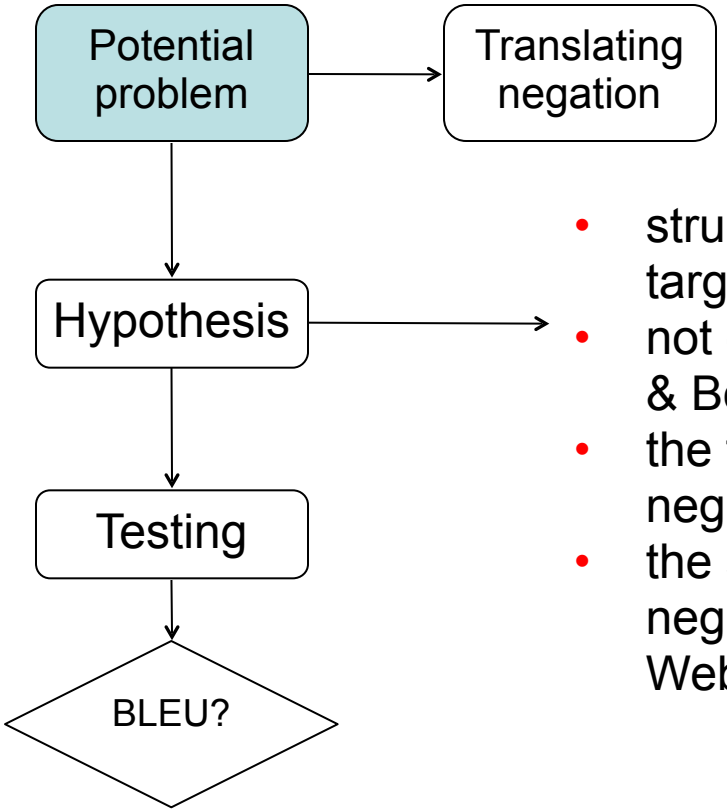
- structural mismatch between source and target language (Collins et al.,2005)
- not enough negative training data (Wetzel & Bond, 2012)
- the translation rules does not contain negation (Baker et al., 2013)

What's exactly wrong with negation?



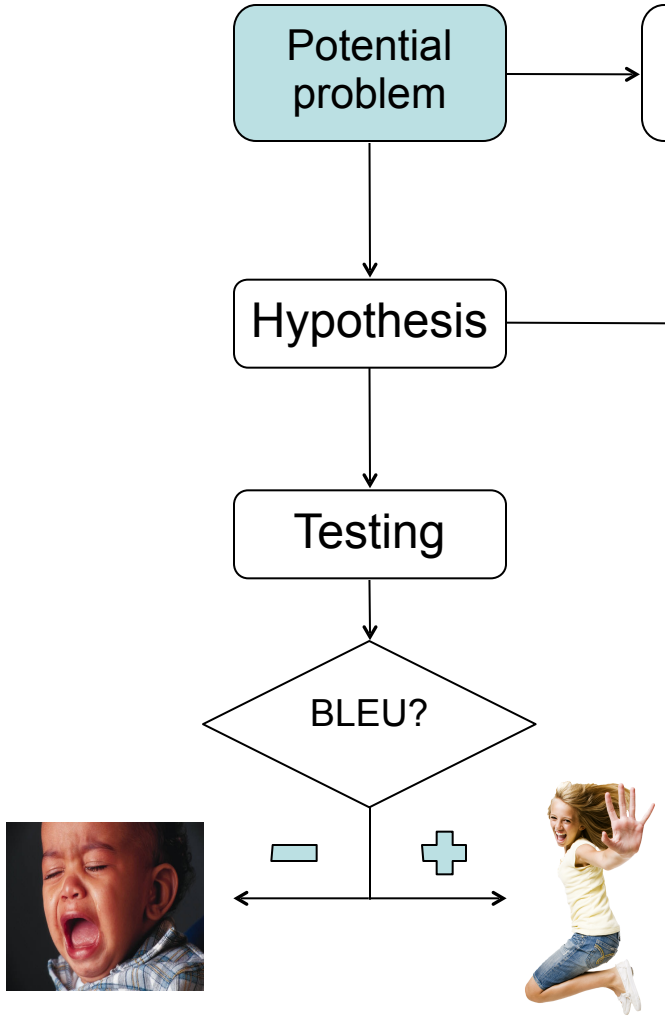
- structural mismatch between source and target language (Collins et al.,2005)
- not enough negative training data (Wetzel & Bond, 2012)
- the translation rules does not contain negation (Baker et al., 2013)
- the scoring function does not contain any negation-related feature (Fancellu & Webber, 2014)

What's exactly wrong with negation?



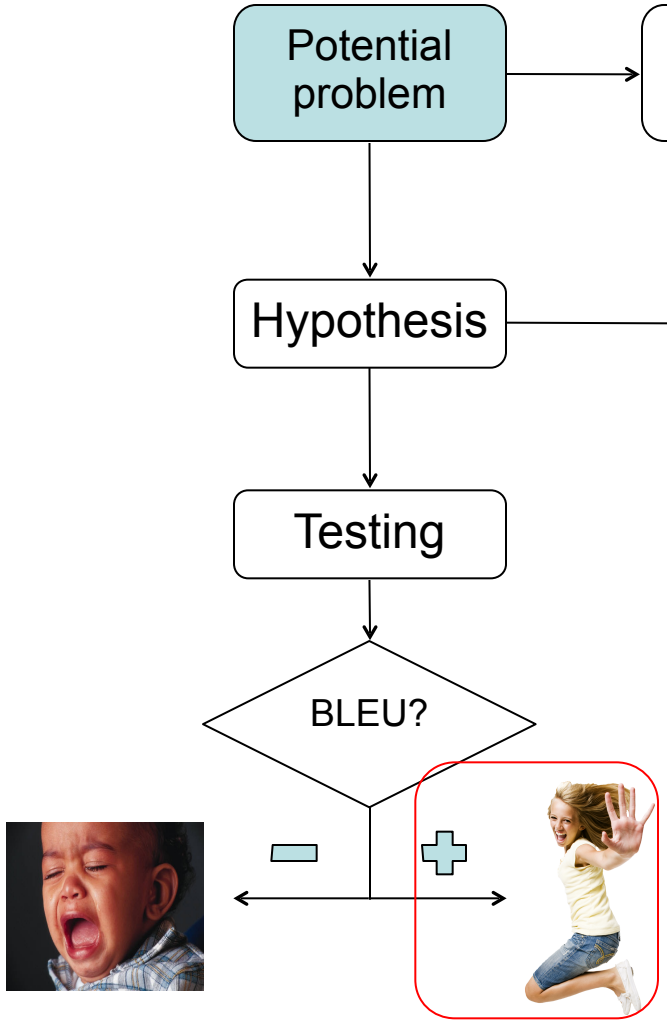
- structural mismatch between source and target language (Collins et al.,2005)
- not enough negative training data (Wetzel & Bond, 2012)
- the translation rules does not contain negation (Baker et al., 2013)
- the scoring function does not contain any negation-related feature (Fancellu & Webber, 2014)

What's exactly wrong with negation?



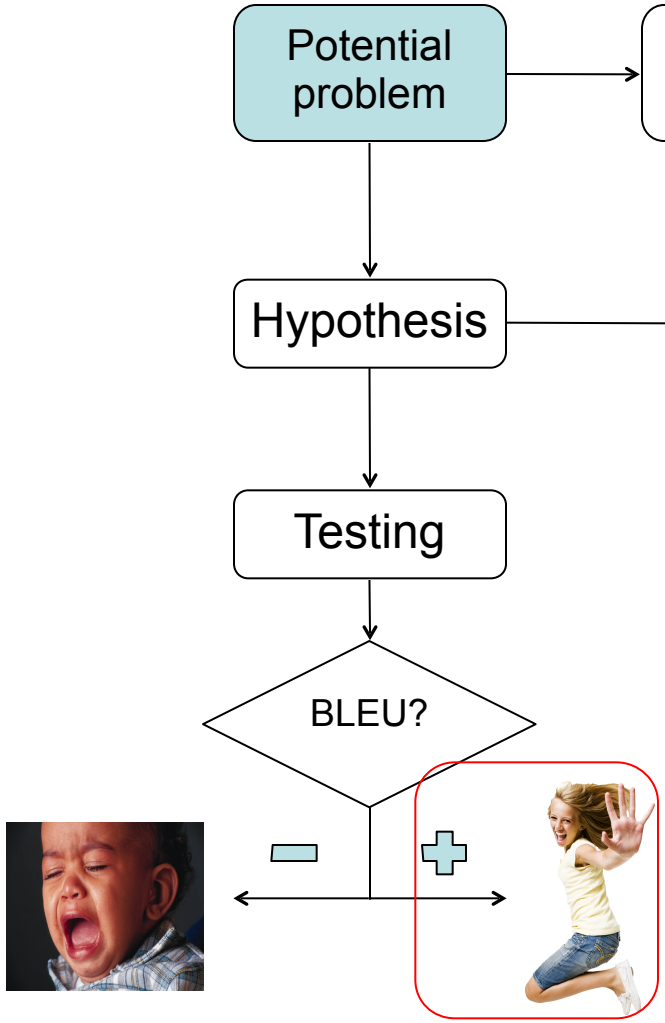
- structural mismatch between source and target language (Collins et al.,2005)
- not enough negative training data (Wetzel & Bond, 2012)
- the translation rules does not contain negation (Baker et al., 2013)
- the scoring function does not contain any negation-related feature (Fancellu & Webber, 2014)

What's exactly wrong with negation?



- structural mismatch between source and target language (Collins et al.,2005)
- not enough negative training data (Wetzel & Bond, 2012)
- the translation rules does not contain negation (Baker et al., 2013)
- the scoring function does not contain any negation-related feature (Fancellu & Webber, 2014)

What's exactly wrong with negation?



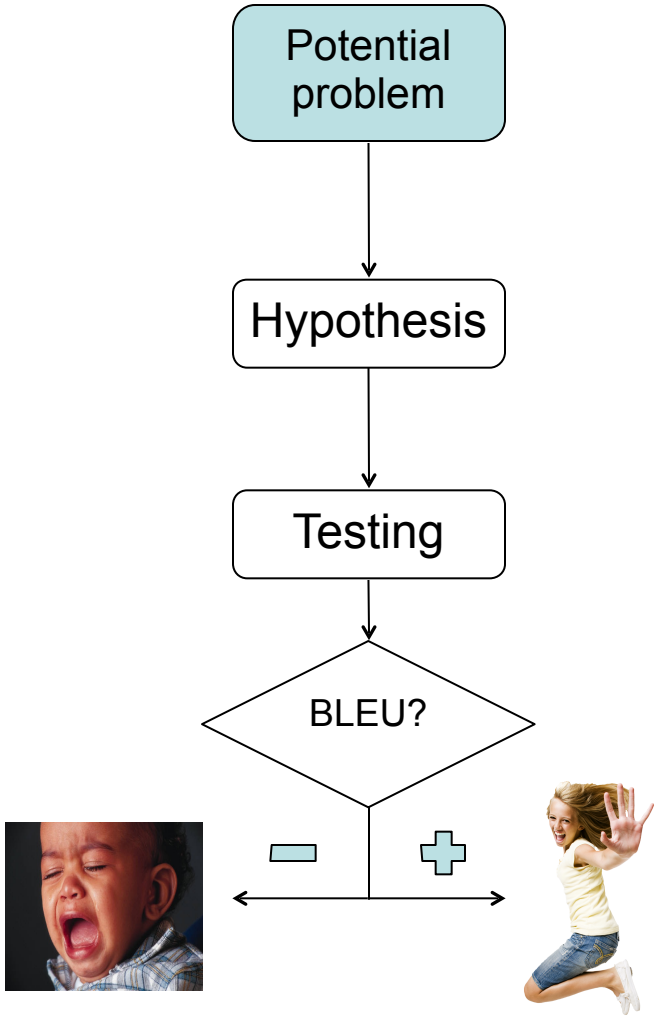
- structural mismatch between so[illegible]and target language (Collins et a[illegible]05)
- not enough negative trai[illegible]data (Wetzel & Bond, 2012)
- the translation ru[illegible]oes not contain negation (Ba[illegible]t al., 2013)
- the scori[illegible]nction does not contain any nega[illegible]elated feature (Fancellu & W[illegible]er, 2014)

NO ERROR ANALYSIS

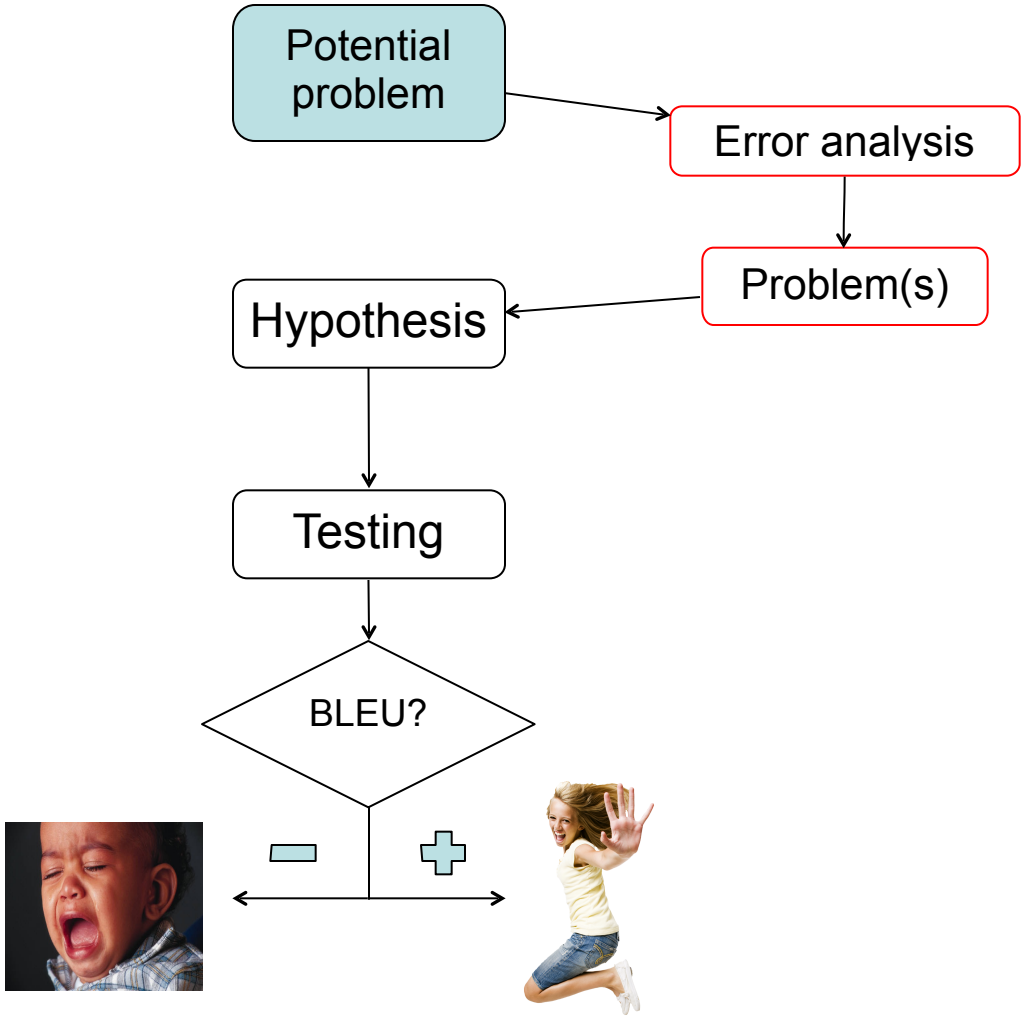
THE UNIVERSITY of EDINBURGH
informatics

Rationale

Rationale



Rationale


Potential problem
Error analysis
Problem(s)
Hypothesis
Testing
BLEU?

THE UNIVERSITY of EDINBURGH
informatics

Contributions

THE UNIVERSITY of EDINBURGH
informatics

Contributions

- Present ongoing work on:

Contributions

- Present ongoing work on:
  - Finding the *causes* of negation-related error during decoding

Contributions

- Present ongoing work on:
  - Finding the *causes* of negation-related error during decoding
  - Highlighting the shortcomings of previous techniques
    - Constrained decoding

Contributions

- Present ongoing work on:
  - Finding the *causes* of negation-related error during decoding
  - Highlighting the shortcomings of previous techniques
    - Constrained decoding
  - Develop an informative way to analyze the translation of negation at each step during decoding
    - **Chart analysis**

Sub-constituents of negation

在 同 一 个 急诊 的 值班 中 ， 我 两 次 没有 发现 病患 得了 盲肠炎 。

During my emergency duty , I have n’t diagnosed a patient with appendicitis twice .

Sub-constituents of negation

在 同 一 个 急诊 的 值班 中 ， 我 两 次 没有 发现 病患 得了 盲肠炎 。

During my emergency duty , I have n't diagnosed a patient with appendicitis twice .

- Cue : the morpheme, word or multi-word unit inherently expressing negation.
  - *im*-possible, breath*less*ness, 不要脸， 不少，…
  - *by no means, save, …*

Sub-constituents of negation

在 同 一 个 急诊 的 值班 中 ， 我 两 次 没有 发现 病患 得了 盲肠炎 。

During my emergency duty , I have n't diagnosed a patient with appendicitis twice .

- Cue : the morpheme, word or multi-word unit inherently expressing negation.
  - *im*-possible, breath*less*ness, 不要脸， 不少，…
  - *by no means, save, …*
- Event : the lexical unit the cue directly refers to

Sub-constituents of negation

在 同 一 个 急诊 的 值班 中 , 我 两 次 没有 发现 病患 得了 盲肠炎 。

During my emergency duty , I have n't diagnosed a patient with appendicitis twice .

- Cue : the morpheme, word or multi-word unit inherently expressing negation.
  - *im*-possible, breath*less*ness, 不要脸， 不少，…
  - *by no means, save, …*
- Event : the lexical unit the cue directly refers to
- Scope: all the elements whose falsity would prove negation to be false.
  - The event is included in the scope

THE UNIVERSITY of EDINBURGH
informatics

What kind of errors?

What kind of errors?

- Manual analysis of the errors involved in translating negation (Fancellu & Webber, 2015 – Ex-Prom @ NAACL ‘15)

What kind of errors?

- Manual analysis of the errors involved in translating negation (Fancellu & Webber, 2015 – Ex-Prom @ NAACL '15)
  - Annotation of the sub-constituents of negation

What kind of errors?

- Manual analysis of the errors involved in translating negation (Fancellu & Webber, 2015 – Ex-Prom @ NAACL '15)
  - Annotation of the sub-constituents of negation
  - HMEANT (Lo & Wu, 2010) to calculate P, R and F1 measure

What kind of errors?

- Manual analysis of the errors involved in translating negation (Fancellu & Webber, 2015 – Ex-Prom @ NAACL '15)
  - Annotation of the sub-constituents of negation
  - HMEANT (Lo & Wu, 2010) to calculate P, R and F1 measure
  - Classification of the errors into **deletion**, **reordering** and **insertion** errors

What kind of errors?

- Manual analysis of the errors involved in translating negation (Fancellu & Webber, 2015 – Ex-Prom @ NAACL '15)
  - Annotation of the sub-constituents of negation
  - HMEANT (Lo & Wu, 2010) to calculate P, R and F1 measure
  - Classification of the errors into **deletion**, **reordering** and **insertion** errors
  - Results:

What kind of errors?

- Manual analysis of the errors involved in translating negation (Fancellu & Webber, 2015 – Ex-Prom @ NAACL '15)
  - Annotation of the sub-constituents of negation
  - HMEANT (Lo & Wu, 2010) to calculate P, R and F1 measure
  - Classification of the errors into **deletion**, **reordering** and **insertion** errors
  - Results:
    - Cue is easiest to translate followed by event and scope difficult

What kind of errors?

- Manual analysis of the errors involved in translating negation (Fancellu & Webber, 2015 – Ex-Prom @ NAACL '15)
  - Annotation of the sub-constituents of negation
  - HMEANT (Lo & Wu, 2010) to calculate P, R and F1 measure
  - Classification of the errors into **deletion**, **reordering** and **insertion** errors
  - Results:
    - Cue is easiest to translate followed by event and scope difficult
    - **Deletion** across all categories

What kind of errors?

- Manual analysis of the errors involved in translating negation (Fancellu & Webber, 2015 – Ex-Prom @ NAACL '15)
  - Annotation of the sub-constituents of negation
  - HMEANT (Lo & Wu, 2010) to calculate P, R and F1 measure
  - Classification of the errors into **deletion**, **reordering** and **insertion** errors
  - Results:
    - Cue is easiest to translate followed by event and scope difficult
    - **Deletion** across all categories
    - **Scope reordering**

THE UNIVERSITY *of* EDINBURGH
informatics

What is the source of these errors?

What is the source of these errors?

- **Rule/phrase Table**: the best translation cannot be generated because its necessary phrases/rules are absent from the search space → **induction** errors

What is the source of these errors?

- **Rule/phrase Table**: the best translation cannot be generated because its necessary phrases/rules are absent from the search space → **induction** errors
- **Search space**: the most probable output is absent from the search space → **search** errors

What is the source of these errors?

- **Rule/phrase Table**: the best translation cannot be generated because its necessary phrases/rules are absent from the search space → **induction** errors
- **Search space**: the most probable output is absent from the search space → **search** errors
- **Model**: the model scores a sub-optimal translation higher than an optimal one → **model** errors

THE UNIVERSITY of EDINBURGH
informatics

Constrained decoding

Constrained decoding

- Tries to reconstruct the reference

Constrained decoding

- Tries to reconstruct the reference
- **Reference reachability** as a proxy to analyze errors during decoding

Constrained decoding

- Tries to reconstruct the reference
- **Reference reachability** as a proxy to analyze errors during decoding
- Implemented as a feature in Moses:
  - 1 if the hypothesis is a sub-string of the reference
  - - inf if the hypothesis is not a sub-string of the reference

THE UNIVERSITY of EDINBURGH
informatics

Constrained Decoding

Constrained Decoding

- If the reference is reconstructed:

Constrained Decoding

- If the reference is reconstructed:
  - **Search vs. model errors** (Wisniewski and Yvon, 2013):
    - if p(*e*) < p(*ê*): **search** error
    - if p(*e*) > p(*ê*): **model** error

**e*: 1-best hypothesis
ê : reconstructed reference

Constrained Decoding

- If the reference is reconstructed:
  - **Search vs. model errors** (Wisniewski and Yvon, 2013):
    - if p(*e*) < p(*ê*): **search** error
    - if p(*e*) > p(*ê*): **model** error

*e: 1-best hypothesis
ê : reconstructed reference

- If the reference can *not* be reconstructed:

Constrained Decoding

- If the reference is reconstructed:
  - **Search vs. model errors** (Wisniewski and Yvon, 2013):
    - if p(*e*) < p(*ê*): **search** error
    - if p(*e*) > p(*ê*): **model** error

**e*: 1-best hypothesis
ê : reconstructed reference

- If the reference can *not* be reconstructed:
  - Increase the *translation option limit* (Auli and Lopez, 2009)
    - if the reference can now be reconstructed → **induction** error

Constrained Decoding

- If the reference is reconstructed:
  - **Search vs. model errors** (Wisniewski and Yvon, 2013):
    - if p(*e*) < p(*ê*): **search** error
    - if p(*e*) > p(*ê*): **model** error

**e*: 1-best hypothesis
ê : reconstructed reference

- If the reference can *not* be reconstructed:
  - Increase the *translation option limit* (Auli and Lopez, 2009)
    - if the reference can now be reconstructed → **induction** error
  - Increase the *cube pruning pop limit*
    - if the reference can now be reconstructed → **search** error

THE UNIVERSITY of EDINBURGH
informatics

Locality issues

Locality issues

- Negation is usually a **local** phenomenon

Locality issues

- Negation is usually a **local** phenomenon

就 拿 住 在 村 东南 一个 小 弯 子 里 的 湾 家人 来 说 吧， 虽然 那 一家 子 的 家长 有点 不要脸， 我们 伟大 的 中 村 不是 照样 会 罩 着 这 一 家 吗？

Locality issues

- Negation is usually a **local** phenomenon

就 拿 住 在 村 东南 一个 小 弯 子 里 的 湾 家人 来 说 吧， 虽然 那 一家 子 的 家长 有点 不要脸， 我们 伟大 的 中 村 不是 照样 会 罩 着 这 一 家 吗？

- If we fail to reconstruct a **whole** reference, it is unclear whether it is because of negation

Locality issues

- Negation is usually a **local** phenomenon

就 拿 住 在 村 东南 一个 小 弯 子 里 的 湾 家人 来 说 吧， 虽然 那 一家 子 的 家长 有点 不要脸， 我们 伟大 的 中 村 不是 照样 会 罩 着 这 一 家 吗？

- If we fail to reconstruct a **whole** reference, it is unclear whether it is because of negation

- Solution: isolate the part containing negation and use them as input to CD

Locality issues

- Negation is usually a **local** phenomenon

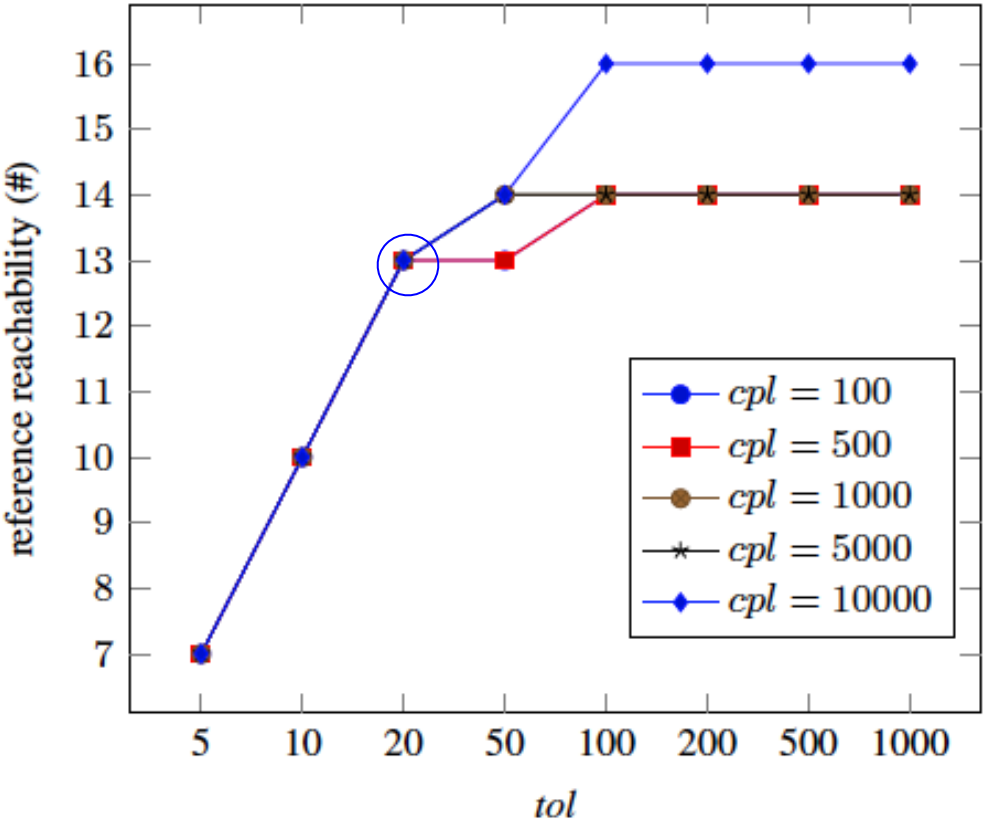
那 一家 子 的 家长 有点 不要脸

the parents of the family are somewhat shameless

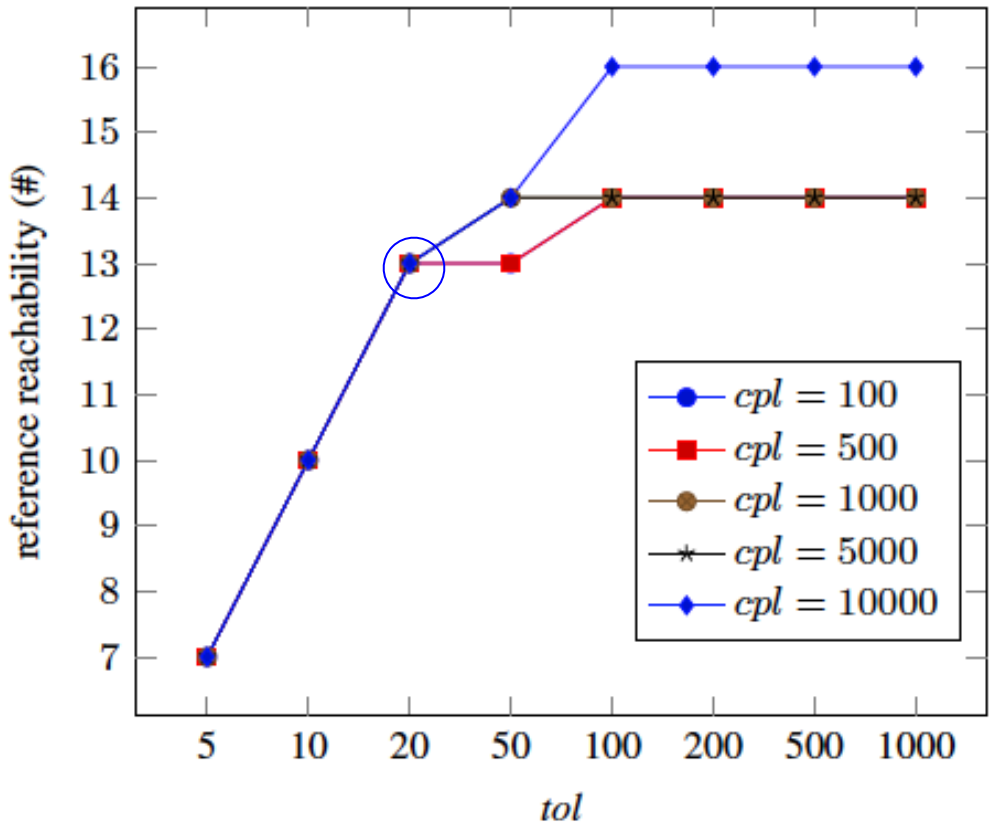
- If we fail to reconstruct a **whole** reference, it is unclear whether it is because of negation

- Solution: isolate the part containing negation and use them as input to CD

Results

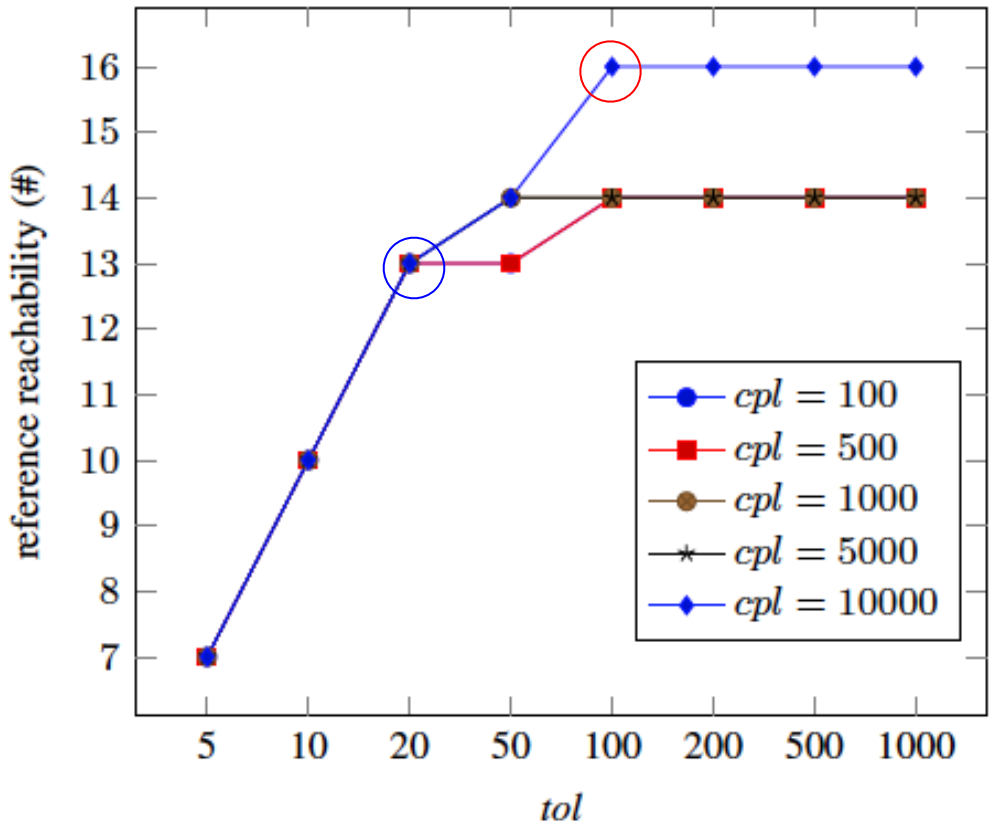

reference reachability (#)
16
15
14
13
12
11
10
9
8
7
5
10
20
50
100
200
500
1000
tol
cpl = 100
cpl = 500
cpl = 1000
cpl = 5000
cpl = 10000

Results



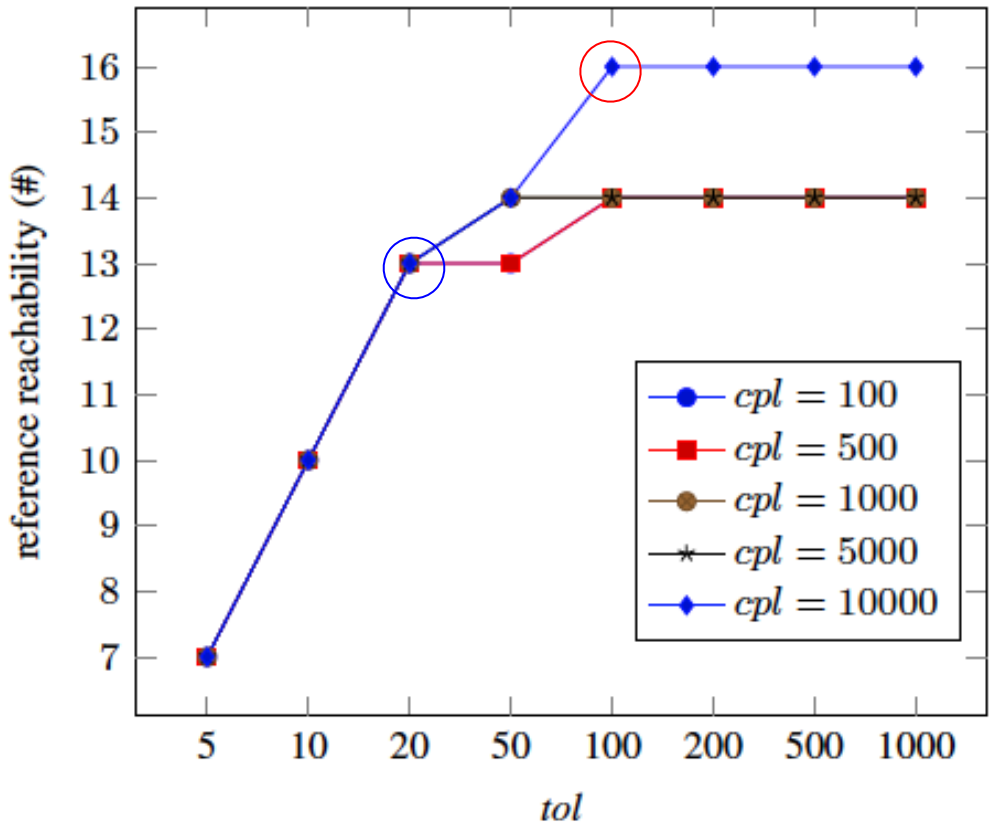
- We could generate max. 16 out of 54 sentences (29%)

Results



- We could generate max. 16 out of 54 sentences (29%)
- Enlarging *translation option limit* and *cube pruning pop limit* leads to a small improvement
  - Just a few **induction/ search** errors

Results



- We could generate max. 16 out of 54 sentences (29%)
- Enlarging *translation option limit* and *cube pruning pop limit* leads to a small improvement
  - Just a few **induction/ search** errors
- p(e) **always** < p(*ê*)
  - **model** errors

THE UNIVERSITY of EDINBURGH
informatics

Discussion

Discussion

- *Ad-interim* conclusion: one should enhance the **model**

Discussion

- *Ad-interim* conclusion: one should enhance the **model**
- **However:**

Discussion

- *Ad-interim* conclusion: one should enhance the **model**
- **However:**
  - We are basing our results on less than a half test sentences
    - ! CD is based only one or a few references vs. virtually infinite ways of translating a sentence

Discussion

- *Ad-interim* conclusion: one should enhance the **model**
- **However:**
  - We are basing our results on less than a half test sentences
    - ! CD is based only one or a few references vs. virtually infinite ways of translating a sentence
  - If **model** errors, which score component is the most responsible?

Discussion

- *Ad-interim* conclusion: one should enhance the **model**
- **However:**
  - We are basing our results on less than a half test sentences
    - ! CD is based only one or a few references vs. virtually infinite ways of translating a sentence
  - If **model** errors, which score component is the most responsible?
  - CD treats decoding as a “black box”

Discussion

- *Ad-interim* conclusion: one should enhance the **model**
- **However:**
  - We are basing our results on less than a half test sentences
    - ! CD is based only one or a few references vs. virtually infinite ways of translating a sentence
  - If **model** errors, which score component is the most responsible?
  - CD treats decoding as a “black box”
  - It is hard to connect CD with **deletion** and **reordering** errors

Chart analysis

- Analysis of each step during decoding
- Access to hypothesis stacks and sub-scores
  - In-depth analysis of **model** errors
- We can understand the causes of **deletion** and **reordering** errors
- We can analyze the translation of cue, event and scope separately
- We can analyze patterns of translation amongst these elements

THE UNIVERSITY of EDINBURGH
informatics

How does it work?

How does it work?

- Input → decoding chart trace

How does it work?

- Input → decoding chart trace
- A good translation of negation needs to meet four conditions:
    1. The cue has to be translated
    2. The event has to be translated
    3. The cue has to refer to the right event
    4. The scope elements should be placed in the correct negation scope

How does it work?

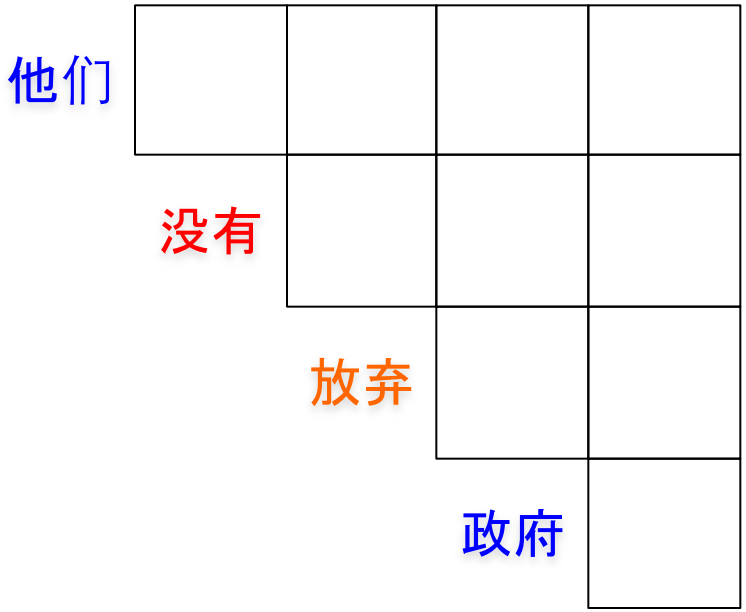
- Input → decoding chart trace
- A good translation of negation needs to meet four conditions:

deletion
1. The cue has to be translated
2. The event has to be translated

reordering
3. The cue has to refer to the right event
4. The scope elements should be placed in the correct negation scope

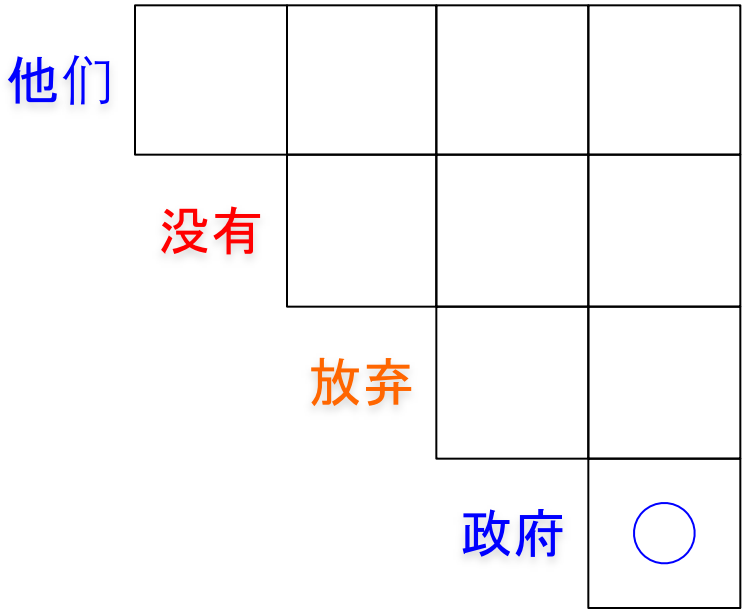
How does it work? – Cont'd

- Assuming we know the elements of negation on the source, the cell has to satisfy a given condition if it cover one or more of those elements



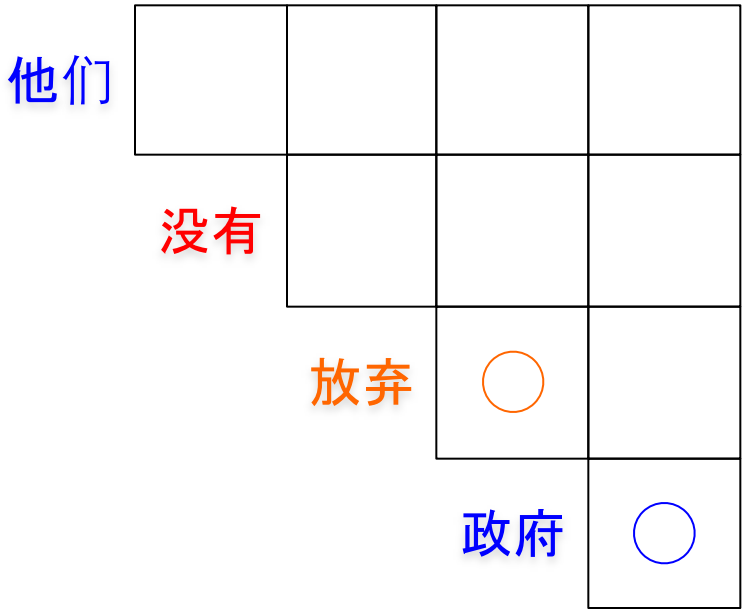
How does it work? – Cont'd

- Assuming we know the elements of negation on the source, the cell has to satisfy a given condition if it cover one or more of those elements



How does it work? – Cont'd

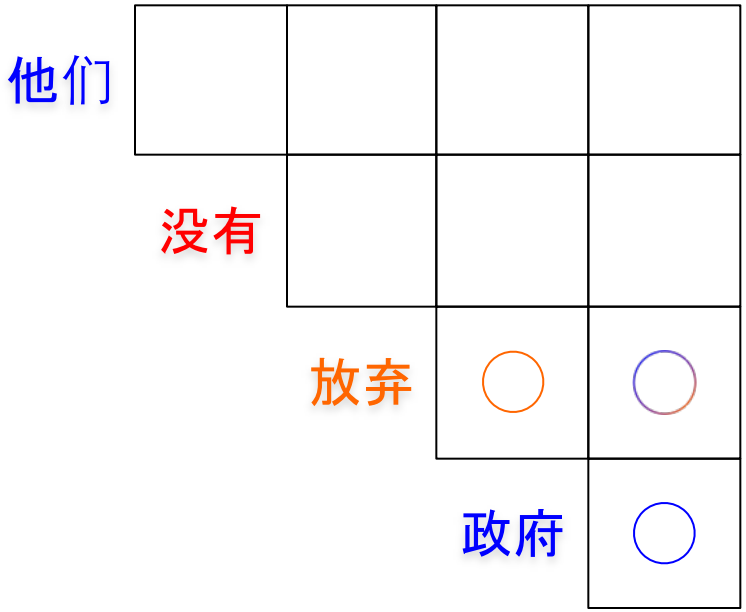
- Assuming we know the elements of negation on the source, the cell has to satisfy a given condition if it cover one or more of those elements



event needs to be translated

How does it work? – Cont'd

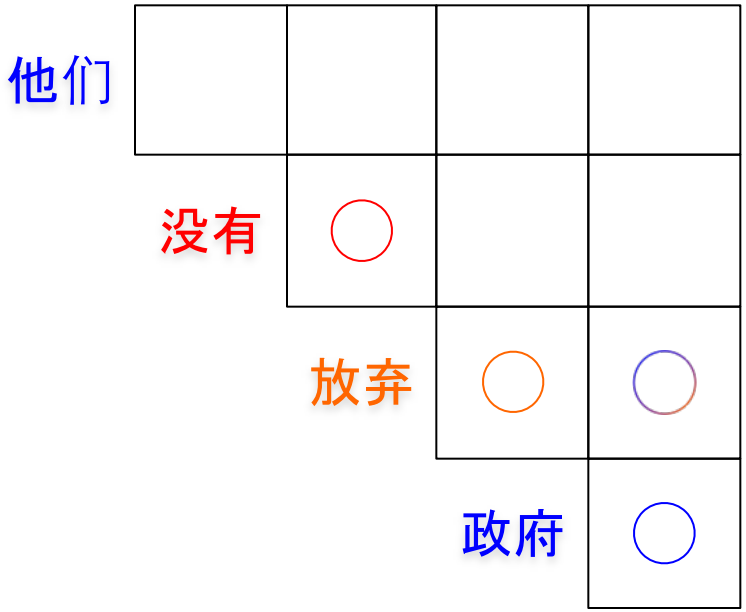
- Assuming we know the elements of negation on the source, the cell has to satisfy a given condition if it cover one or more of those elements



scope element attached to the right
event

How does it work? – Cont'd

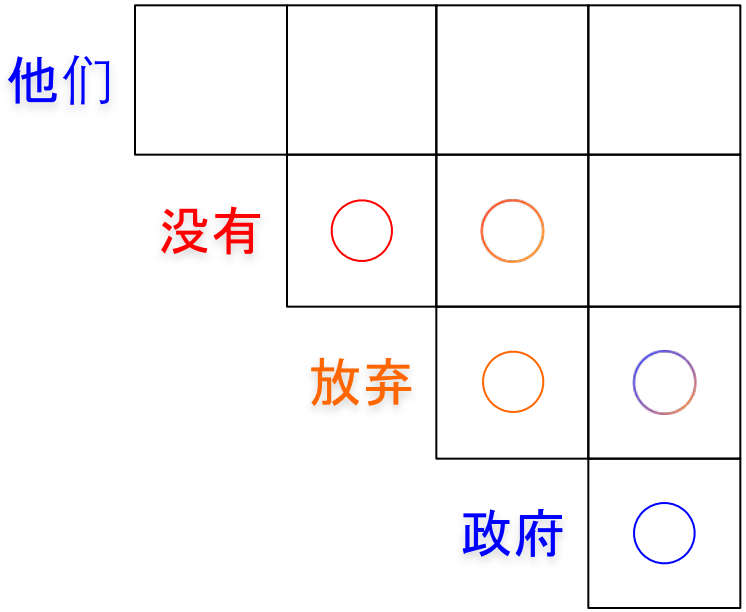
- Assuming we know the elements of negation on the source, the cell has to satisfy a given condition if it cover one or more of those elements



cue needs to be translated

How does it work? – Cont'd

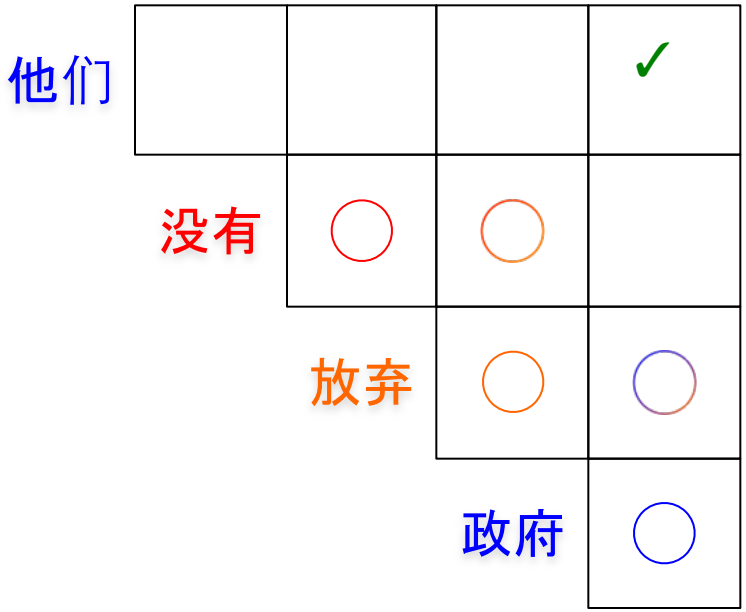
- Assuming we know the elements of negation on the source, the cell has to satisfy a given condition if it cover one or more of those elements



cue should refer to the right event

How does it work? – Cont'd

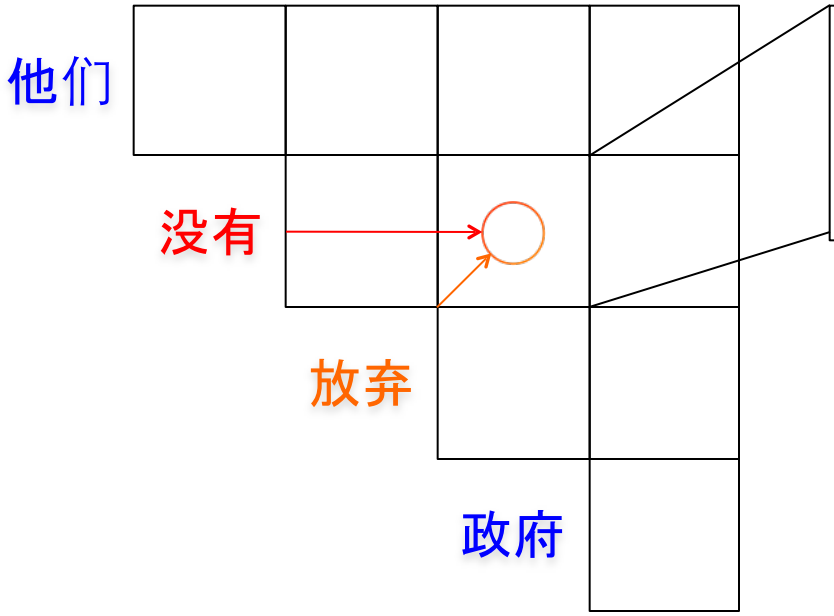
- Assuming we know the elements of negation on the source, the cell has to satisfy a given condition if it cover one or more of those elements



All elements should be translated and should correctly related to each other

Stack analysis – model errors

- Analysis whether a component is more responsible for **model** errors



1. gave up | p(e|f) p(f|e) p(LM) $p_{lex.}$ … ✖
2. not | p(e|f) p(f|e) p(LM) $p_{lex.}$ …

[…] ✓

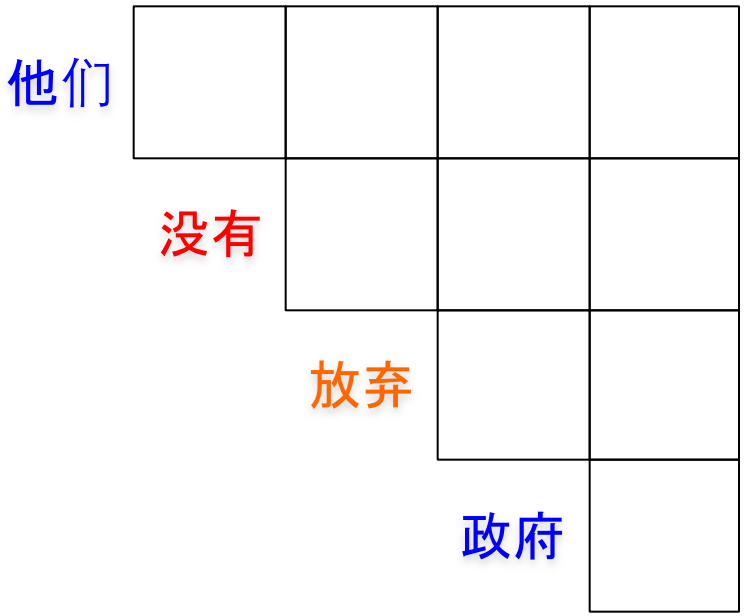
10: did not give up | p(e|f) p(f|e) p(LM) $p_{lex.}$ …

10 meets all conditions, 1 does not

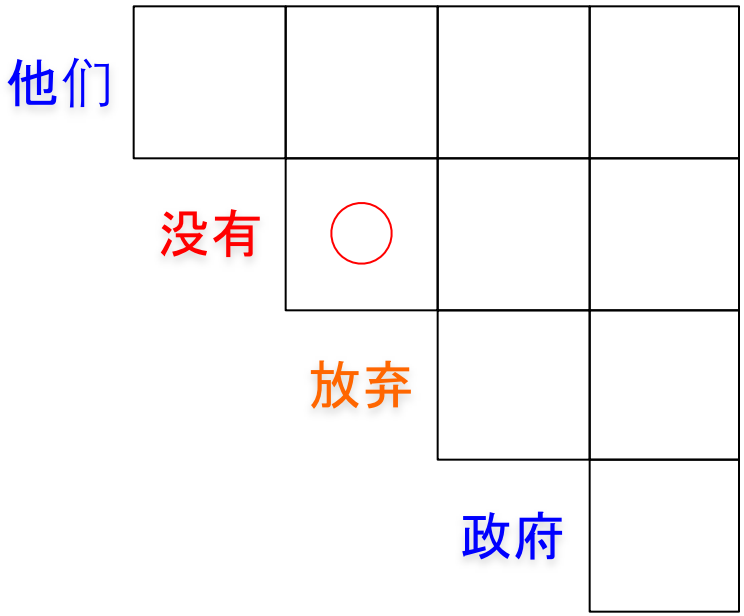
1: p(e|f) p(f|e) p(LM) p_{lex}(e|f) p_{lex}(e|f)

10: p(e|f) p(f|e) p(LM) p_{lex}(e|f) p_{lex}(e|f)

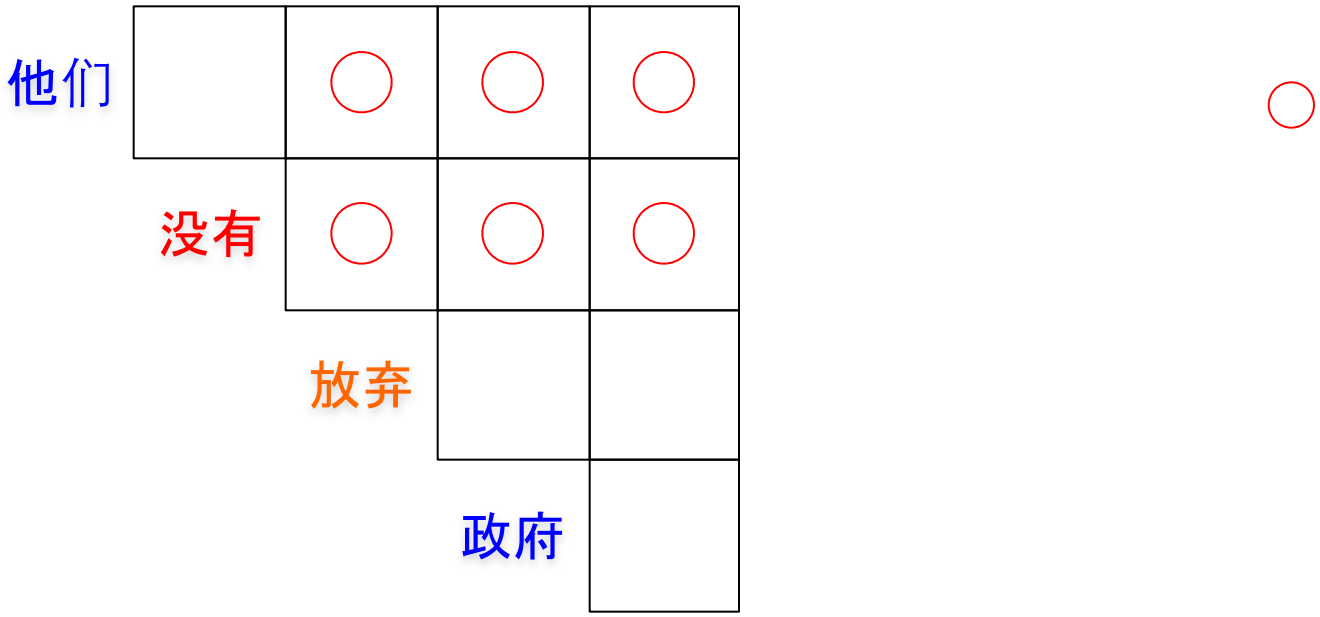
Stack analysis – search/induction errors



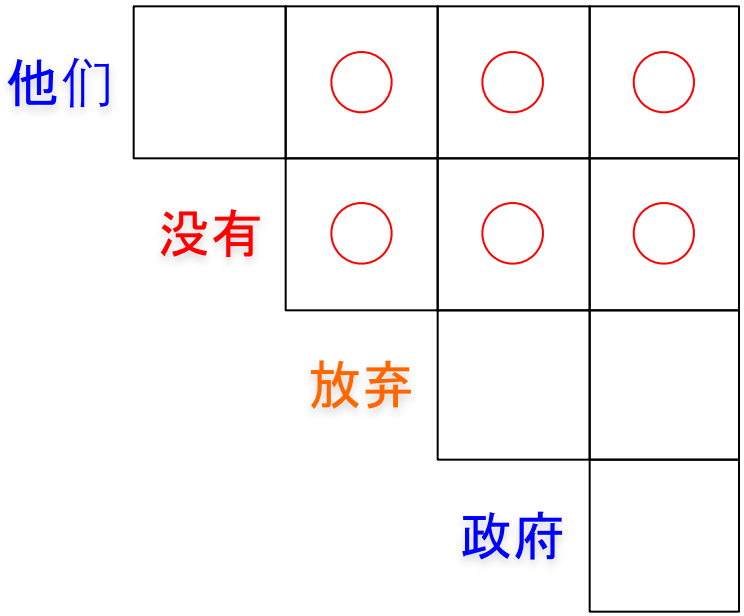
Stack analysis – search/induction errors



Stack analysis – search/induction errors

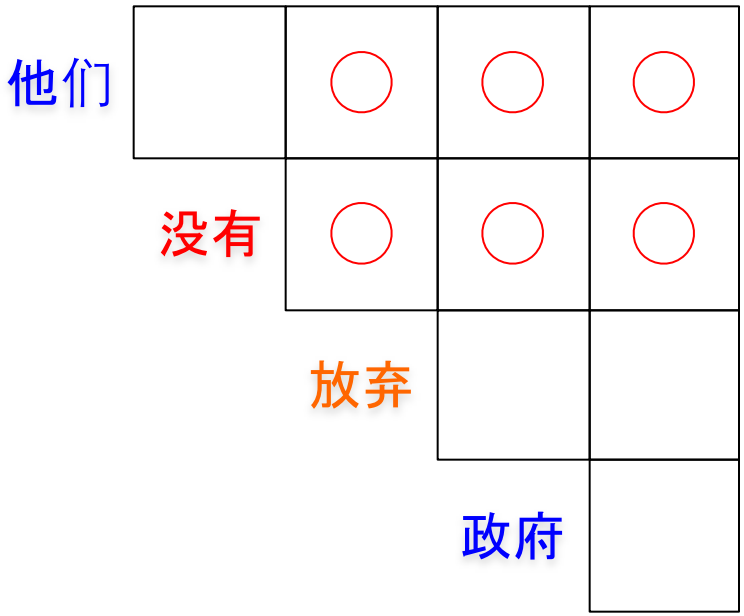


Stack analysis – search/induction errors



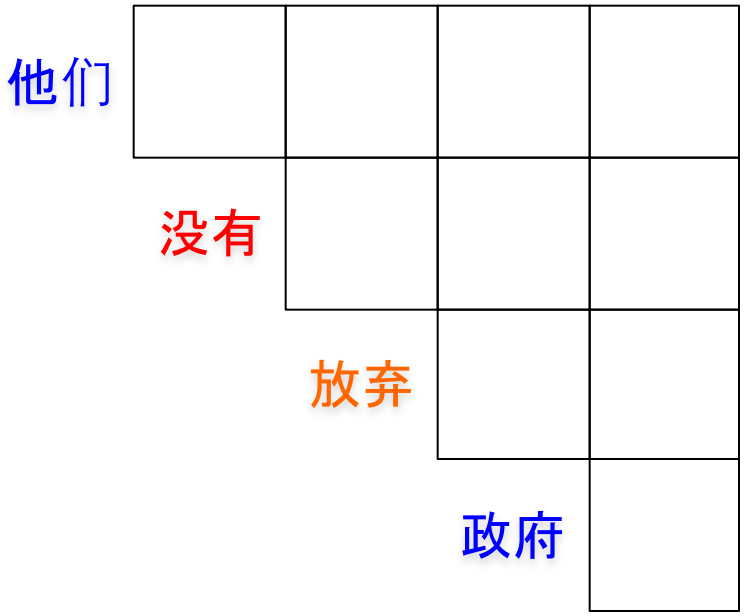
- cue has to be translated in all cells marked with ○

Stack analysis – search/induction errors

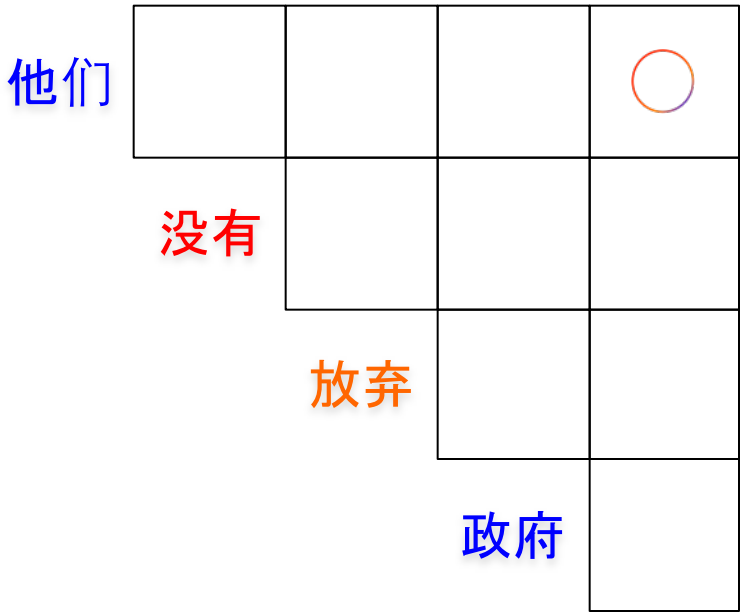


- cue has to be translated in all cells marked with ◯
- If no cue is found in any of these cells:
  - Modify *translation option limit* and *cube pruning pop limit* to assess the presence of **search** and **model** errors
- Same applies to the other two elements

Stack analysis – others/ongoing

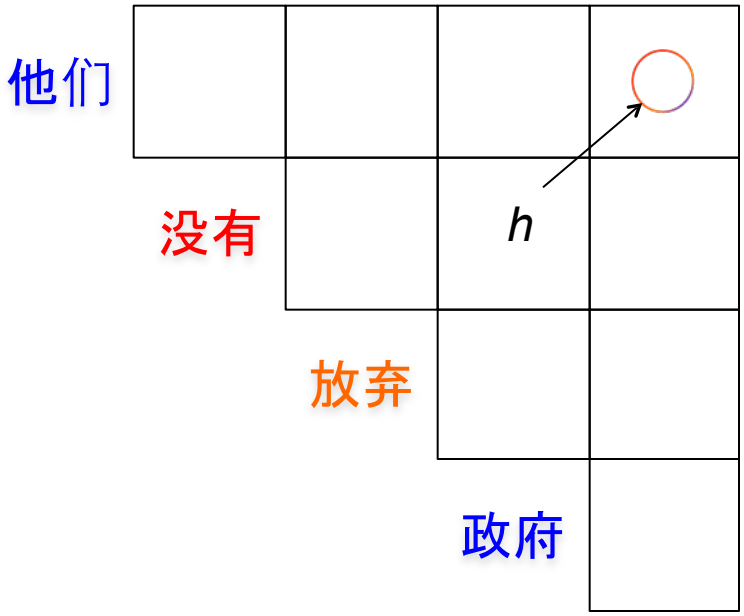


Stack analysis – others/ongoing



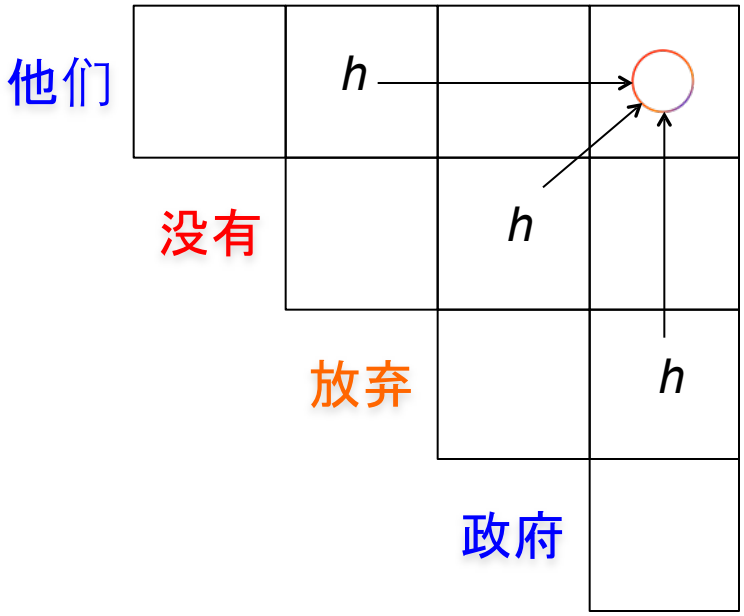
- Rule trace to study negation element combinatory tendencies

Stack analysis – others/ongoing



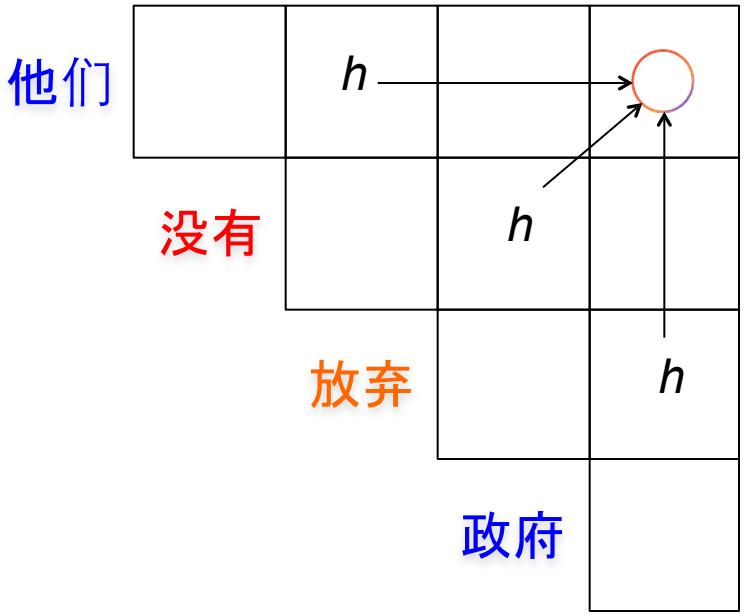
- Rule trace to study negation element combinatory tendencies
- Is cue translated along side the event?

Stack analysis – others/ongoing



- Rule trace to study negation element combinatory tendencies
- Is cue translated along side the event?
- Is cue and event translated separately and combined together via glue rules?

Stack analysis – others/ongoing



- Rule trace to study negation element combinatory tendencies
- Is cue translated along side the event?
- Is cue and event translated separately and combined together via glue rules?
- What about event and scope?

Negation detection

- Source → annotations from manual error analysis
- Target?

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]

Negation detection

- Source → annotations from manual error analysis
- Target?

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]

Negation detection

- Source → annotations from manual error analysis
- Target?

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]

Negation detection

- Source → annotations from manual error analysis
- Target?

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]

Negation detection

- Source → annotations from manual error analysis
- Target?

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]

cue

Not
No
Neither
Impossible
By no means
[…]

*CRF (F1 > 90%)

Negation detection

- Source → annotations from manual error analysis
- Target?

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]

cue

Not
No
Neither
Impossible
By no means
[…]

*CRF (F1 > 90%)

Negation detection

- Source → annotations from manual error analysis
- Target?

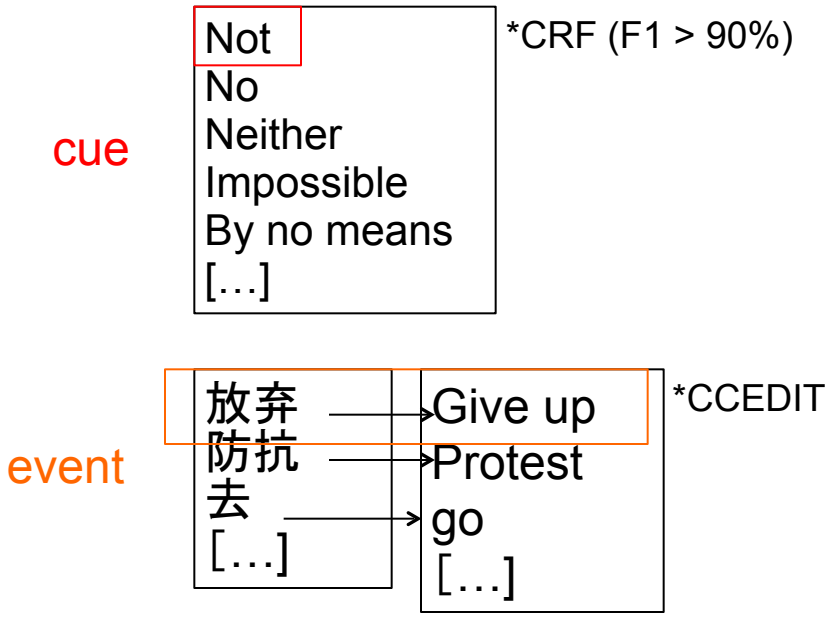
??

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]



Negation detection

- Source → annotations from manual error analysis
- Target?

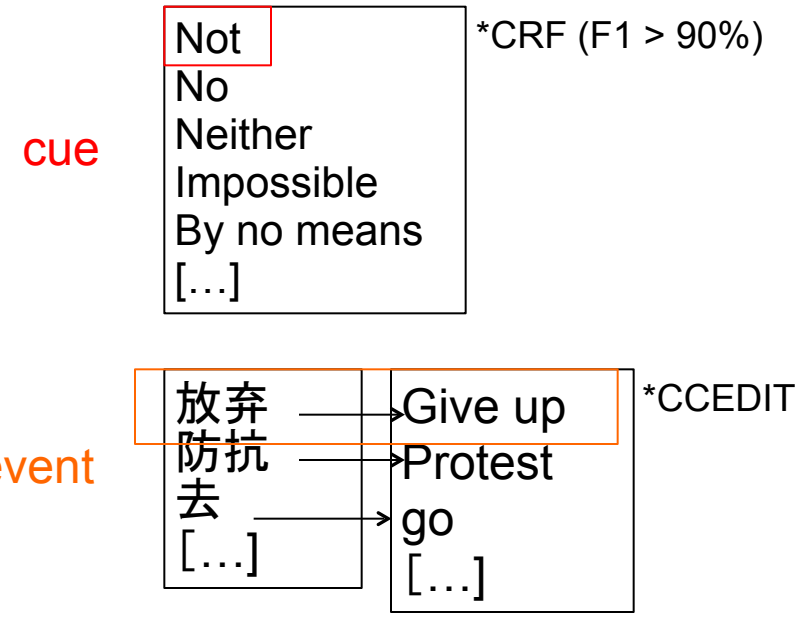
??

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]



Negation detection

- Source → annotations from manual error analysis
- Target?

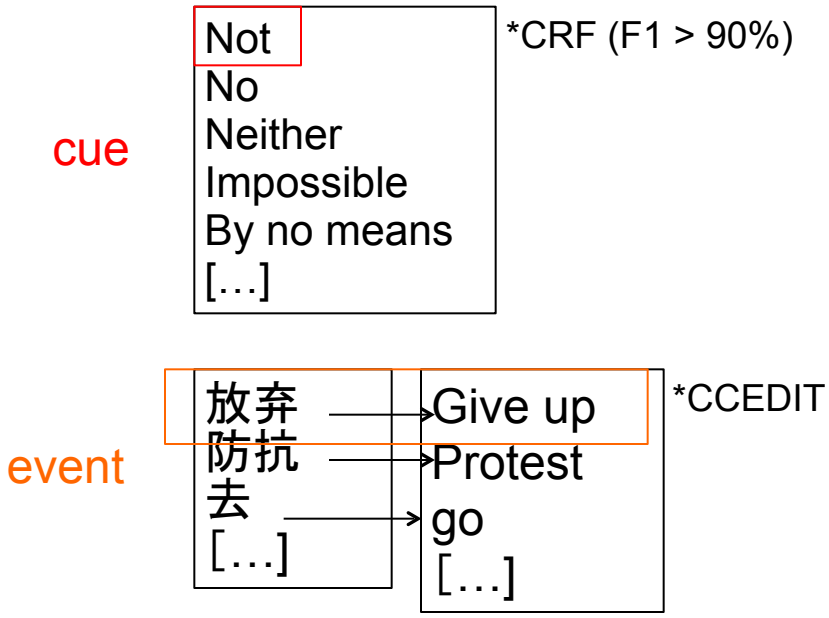
??

1. gave up ||| […]
2. not ||| […]

10: did not give up ||| […]

[…]

25: he did not give up ||| […]



- Better approach: **paraphrase + automatic negation detection** (see Future Work)

System and initial results

- System:
  - Zh → En HIERO; 54 sentences containing negation (from the *manual* error analysis)
- Results:
  - Errors related to the translation of the cue
  - The cue is *never* absent from the chart of *any* sentence
    - no **search** or **induction error**
  - Analysis of the model sub-scores:
    - **Indirect probabilities** (translation and lexical) are responsible for > 60% of bad-ranking
    - **LM** only 25%

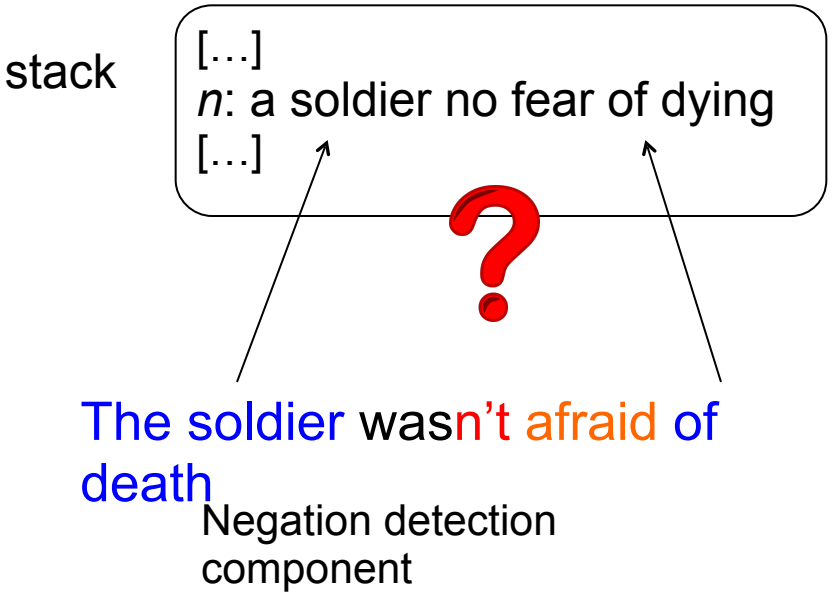
Conclusion

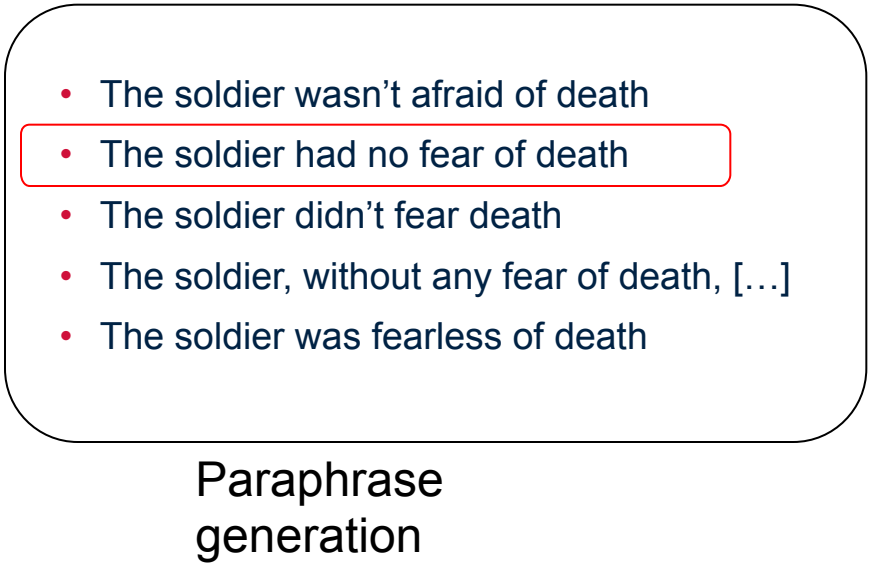
- Translating negation is problematic
- Previous error detection techniques do not offer an in-depth analysis
- A chart analysis offers a better insight in the decoding process

	Model	Search	Induction
Cue	Y	N	N
Event	…	…	…
Scope	…	…	…

Future Work

- Negation detection in the target hypothesis
- No **list!** How to leverage a **reference** translation?





THE UNIVERSITY *of* EDINBURGH
informatics

Thank you!