Definite noun phrases in statistical machine translation into Scandinavian languages

Sara Stymne

Linköping University, Sweden
Xerox Research Centre Europe, France

EAMT, Leuven, Belgium
May 31, 2011

Task

- Improve translation into Danish, Swedish, and Norwegian with a focus on definiteness

Task

- Improve translation into Danish, Swedish, and Norwegian with a focus on definiteness
- Scandinavian languages have two ways to express definiteness
- Leads to errors:
  - Spurious definite articles
  - Wrong form of nouns

Task

- Improve translation into Danish, Swedish, and Norwegian with a focus on definiteness
- Scandinavian languages have two ways to express definiteness
- Leads to errors:
  - Spurious definite articles
  - Wrong form of nouns

- Simple pre-processing solution – that works!

Outline

1 Definiteness

2 Pre-processing

3 Suggested approach

4 Experiments

5 Conclusion

Definiteness – Danish

NP type	**Danish**	**English**
sg, -mod	hund**en**	**the** dog
sg, +mod	**den** sorte hund	**the** black dog

Definiteness – Danish

NP type	**Danish**	**English**
sg, -mod	hund**en**	**the** dog
	***den** hund(**en**)	
sg, +mod	**den** sorte hund	**the** black dog
	***den** sorte hund**en**	
	*sorte hund**en**	

Definiteness – Danish

NP type	**Danish**	**English**
sg, -mod	hund**en**	**the** dog
	***den** hund(**en**)	
sg, +mod	**den** sorte hund	**the** black dog
	***den** sorte hund**en**	
	*sorte hund**en**	
pl, -mod	hunde**ne**	**the** dogs
pl, +mod	**de** sorte hunde	**the** black dogs

Definiteness – Swedish/Norwegian

NP type	**Swedish**	**English**
sg, -mod	hund**en**	**the** dog
sg, +mod	**den** svarta hund**en**	**the** black dog

Definiteness – Swedish/Norwegian

NP type	**Swedish**	**English**
sg, -mod	hund**en**	**the** dog
	***den** hund(**en**)	
sg, +mod	**den** svarta hund**en**	**the** black dog
	***den** svarta hund	
	*svarta hund**en**	

Definiteness – Swedish/Norwegian

NP type	**Swedish**	**English**
sg, -mod	hund**en**	**the** dog
	***den** hund(**en**)	
sg, +mod	**den** svarta hund**en**	**the** black dog
	***den** svarta hund	
	*svarta hund**en**	
pl, -mod	hundar**na**	**the** dogs
pl, +mod	**de** svarta hundar**na**	**the** black dogs

Other uses of definite suffixes

- Demonstratives (SV)
  - This dog
  - Den (här) hund**en**, denna hund

Other uses of definite suffixes

- Demonstratives (SV)
  - This dog
  - Den (här) hund**en**, denna hund
- Possesives (NO)
  - My black dog
  - min svarte hund, **den** svarte hund**en** min

Other uses of definite suffixes

- Demonstratives (SV)
  - This dog
  - Den (här) hund**en**, denna hund
- Possesives (NO)
  - My black dog
  - min svarte hund, **den** svarte hund**en** min
- With relative clauses (DA, SV, NO)
  - The dog that barked is nice
  - **Den** hund som skällde är snäll
  - The dog, which barked, is nice
  - Hund**en** som skällde är snäll

Outline

1 Definiteness

2 Pre-processing

3 Suggested approach

4 Experiments

5 Conclusion

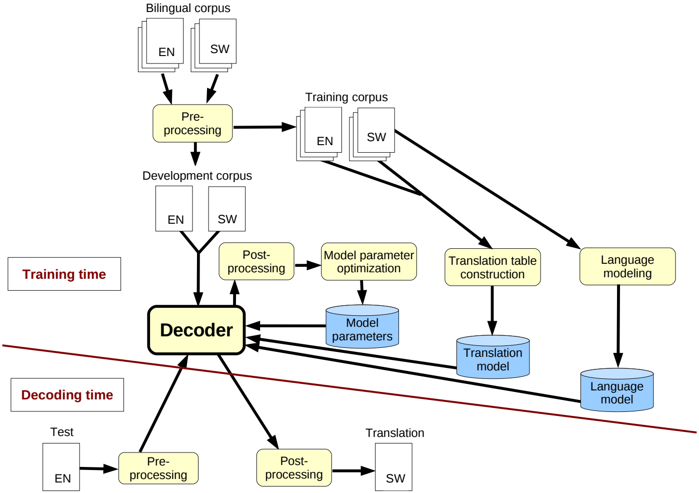
Pre-processing

- Change the source to become more common to the target
- Change the target to become more common to the source (needs post-processing)
- Commonly used, for instance for:
  - Tokenization, lower-casing, etc
  - Compounds
  - Reordering
  - . . .

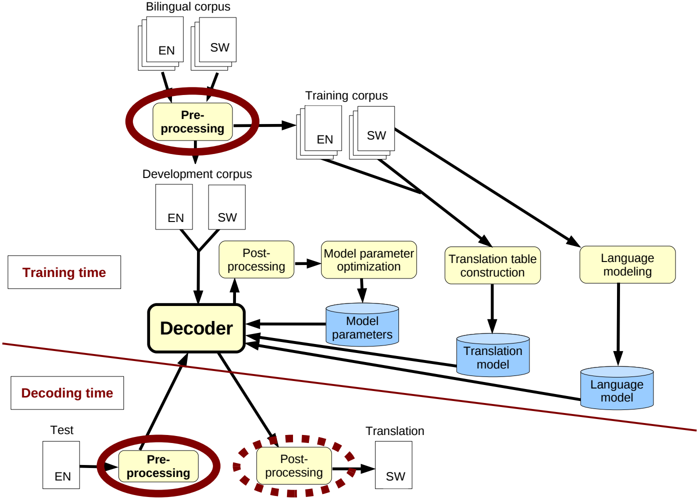
Pre-processing

- Change the source to become more common to the target
- Change the target to become more common to the source (needs post-processing)
- Commonly used, for instance for:
  - Tokenization, lower-casing, etc
  - Compounds
  - Reordering
  - . . .
  - Definiteness, En–Da (Stymne, 2009)

System integration I



System integration II



Applications

- Source side processing:
  - **English** ⇒ Danish, Swedish, Norwegian
  - **Italian** ⇒ Danish
- Target side processing:
  - English ⇒ **Swedish**

Outline

1 Definiteness

2 Pre-processing

3 Suggested approach

4 Experiments

5 Conclusion

Source side processing

- Pre-processing the source, to look like the target with respect to definiteness:
  - Identify definite NPs
  - Apply suitable operators:
    - remove-DEF
    - add-DEFSUFFIX
    - move-ADJ

Identification of definite NPs

- English
- `the (ADJ∨NUM)* NOUN+`
- Italian
- `il/lo (ADJ∨NUM)* NOUN+ ADJ*`

English–Danish

- Non-modified: remove-DEF, add-DEFSUFFIX
- Modified: none

the	commission	
ART	NOUN	
	commission#the	
the	member	states
ART	NOUN	NOUN
	member	states#the
the	71	countries
ART	NUM	NOUN
the	71	countries

Italian–Danish

- Non-modified: remove-DEF, add-DEFSUFFIX
- Modified: move-ADJ

i (il)	livelli	
ART	NOUN	
	livelli-DEF	
nei (in il)	fondi	strutturali
APPRART	NOUN	ADJ
in il	strutturali	fondi

English–SV/NO – strategy 1

- Non-modified: remove-DEF, add-DEFSUFFIX
- Modified: add-DEFSUFFIX

the	commission	
ART	NOUN	
	commission#the	
the	member	states
ART	NOUN	NOUN
	member	states#the
the	71	countries
ART	NUM	NOUN
the	71	countries#the

English–SV/NO – strategy 2

- Non-modified: remove-DEF, ~~add-DEFSUFFIX~~
- Modified: ~~add-DEFSUFFIX~~

the	commission	
ART	NOUN	
	commission	
the	member	states
ART	NOUN	NOUN
	member	states
the	71	countries
ART	NUM	NOUN
the	71	countries

Target side processing

- Pre-processing the target, to look like the source with respect to definiteness:
  - Identify bare definite NPs
    - `¬(ADJ∨NUM∨ART) NOUN.DEF`
  - Add dummy definite articles

English–SV – Target processing

- Non-modified: Add article
- Modified: nothing

	medlemsstaterna	
	NOUN.DEF	
DEF	medlemsstaterna	
the	member states	
det	svåra	beslutet
ART	ADJ	NOUN.DEF
det	svåra	beslutet
the	hard	decision

Outline

1. Definiteness
2. Pre-processing
3. Suggested approach
4. Experiments
5. Conclusion

Decoders

- Moses
  - standard phrase-based SMT
- Matrax
  - phrase-based SMT
  - handles noncontiguous phrases:

Decoders

- Moses
  - standard phrase-based SMT
- Matrax
  - phrase-based SMT
  - handles noncontiguous phrases:

    does not **jeopardize** the effort

    bringer ikke indsatsen **i fare**

Corpora

- Automotive manuals
- Based on translation-memory data
  - English–Danish, 170k sentences
  - English–Swedish, 330k sentences
  - English–Norwegian, 400k sentences
- Europarl
  - Italian–Danish, 100k sentences
  - English–Danish, 700k sentences
  - English–Swedish, 700k sentences

Results – Danish

Languages	**System**	**Bleu**	**NIST**
En-Da – auto	Baseline	70.91	8.8816
	DEF-proc	76.35+	9.3629+
En-Da – auto+P+CS	Baseline	74.09	9.2328
	DEF-proc	76.17+	9.4342+
En-Da – EP	Baseline	19.01	5.6373
	DEF-proc	23.22+	6.1009+

Results – Danish

Languages	System	Bleu	NIST
En-Da – auto	Baseline	70.91	8.8816
	DEF-proc	76.35+	9.3629+
En-Da – auto+P+CS	Baseline	74.09	9.2328
	DEF-proc	76.17+	9.4342+
En-Da – EP	Baseline	19.01	5.6373
	DEF-proc	23.22+	6.1009+
It-Da – EP	Baseline	10.54	4.3924
	DEF-proc	12.04+	4.5754+

Results – Swedish

Languages	System	Bleu	NIST
En-Sv – auto	Baseline	61.20	9.7934
	DEF-proc1	58.84-	9.4898-
	DEF-proc2	62.05+	9.9129+
En-Sv – EP+P	Baseline	21.63	6.1085
	DEF-proc2	22.03+	6.1778+

Results – Swedish

Languages	System	Bleu	NIST
En-Sv – auto	Baseline	61.20	9.7934
	DEF-proc1	58.84-	9.4898-
	DEF-proc2	62.05+	9.9129+
En-Sv – EP+P	Baseline	21.63	6.1085
	DEF-proc2	22.03+	6.1778+
	Target-proc	21.31-	6.1018

Results – Norweigan

Languages	System	Bleu	NIST
En-No	Baseline	58.57	8.8846
	DEF-proc1	56.59-	8.6943-
	DEF-proc2	59.08	8.9092

Tentative error analysis

- English–Swedish Europarl
- Manual analysis of errors in 50 sentences
- Fewest total errors for target proc.
- Source side: more wrong or extra function words
- Definite processing: fewer word order and punctuation errors
- Approximately the same number of definiteness errors in all systems:
  - Source side: only wrong form of nouns
  - Other systems: mixed error types, e.g. spurious articles

Example, Italian–Danish

Src: Non pensa che dovremmo ormai esplorare nuovi modi per affrontare il problema delle nostre relazioni con la Birmania?

Ref: Finder De ikke, at vi bör se på andre måder, hvorpå vi kan tackle problemet med vores relationer i Burma?

Baseline: Tänker ikke at vi bör efterhånden resterende tid nye måder fat af vores forbindelser med den Burma?

DEF-proc: Tänker ikke at vi bör overveje nye måder nu af vores forbindelser med Burma tackle problemet?

Example, Italian–Danish

Src: Non pensa che dovremmo ormai esplorare nuovi modi per affrontare il problema delle nostre relazioni con la Birmania?

Ref: Finder De ikke, at vi bör se på andre måder, hvorpå vi kan tackle problemet med vores relationer i Burma?

Baseline: Tänker ikke at vi bör efterhånden resterende tid nye måder fat af vores forbindelser med den Burma?

DEF-proc: Tänker ikke at vi bör overveje nye måder nu af vores forbindelser med Burma tackle problemet?

Example, English–Swedish

Src:	Men who commit murders rarely receive long prison sentences . . .
Ref:	Männen som utför morden får sällan långa fängelsestraff . . .
Baseline:	De män som begår morden sällan få långa fängelsestraff . . .
DEF-proc2:	Män som begår mord sällan få långa fängelsestraff . . .
Target-proc:	De män som begår morden sällan erhåller långa fängelsestraff . . .

Example, English–Swedish

Src:	Men who commit murders rarely receive long prison sentences . . .
Ref:	Männen som utför morden får sällan långa fängelsestraff . . .
Baseline:	De män som begår morden sällan få långa fängelsestraff . . .
DEF-proc2:	Män som begår mord sällan få långa fängelsestraff . . .
Target-proc:	De män som begår morden sällan erhåller långa fängelsestraff . . .

Example, English–Danish

Src: The majority of the women will be travelling to a conference of members of parliament in Berlin.

Ref: Hovedparten af kvinderne skal af sted til en konference for parlamentsmedlemmer i Berlin.

Baseline: Størstedelen af de kvinder bliver til en konference for parlamentsmedlemmer rejser i Berlin.

DEF-proc: Flertallet af kvinderne bliver rejser til en konference af medlemmer af parlamentet i Berlin.

Example, English–Danish

Src: The majority of the women will be travelling to a conference of members of parliament in Berlin.

Ref: Hovedparten af kvinderne skal af sted til en konference for parlamentsmedlemmer i Berlin.

Baseline: Størstedelen af de kvinder bliver til en konference for parlamentsmedlemmer rejser i Berlin.

DEF-proc: Flertallet af kvinderne bliver rejser til en konference af medlemmer af parlamentet i Berlin.

Example, English–Danish

Src: The majority of the women will be travelling to a conference of members of parliament in Berlin.

Ref: Hovedparten af kvinderne skal af sted til en konference for parlamentsmedlemmer i Berlin.

Baseline: Størstedelen af de kvinder bliver til en konference for parlamentsmedlemmer rejser i Berlin.

DEF-proc: Flertallet af kvinderne bliver rejser til en konference af medlemmer af parlamentet i Berlin.

Outline

1. Definiteness
2. Pre-processing
3. Suggested approach
4. Experiments
5. Conclusion

Conclusions

- Special treatment of definiteness is useful
- Source side preprocessing – positive results for a variety of language pairs and datasets
- Target side processing – no positive results
- Careful modification of preprocessing data needed for new language pairs

Future work

- Thorough error analysis
- More elaborate preprocessing strategy, that takes more constructions into account
- Other language pairs
- Present a lattice to decoder

Thank you for your attention!

Tak for opmærksomheden!

Tack för uppmärksamheten!