

Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål

Kevin Unhammer¹ Trond Trosterud²

¹Department of Linguistics
University of Bergen
Bergen, Norway
kun041@student.uib.no

²Department of Linguistics
University of Tromsø
Tromsø, Norway
trond.trosterud@uit.no

2nd November 2009

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Outline of talk

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Reuse of Free
Resources in
Nynorsk↔Bokmål
MT

Kevin Unhammer,
Trond Trosterud

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

The Norwegian language(s)

- ▶ A lot of dialectal variation
- ▶ Two written variants:
 - ▶ Bokmål
 - ▶ Based on Danish and the Dano-Norwegian koiné of the major cities in the 1800's
 - ▶ Nynorsk
 - ▶ Based on the spoken dialects of Norway, standardised by linguist Ivar Aasen in the late 1800's
- ▶ Nynorsk used by around 12% of the population
- ▶ “Language-friendly” politics: Both standards are officially recognised and both are taught in school from age 12 and up
- ▶ Both Nynorsk and Bokmål allow quite a lot of variation, with some choices being considered more “radical” or “conservative” than others

Free, Open Source Norwegian language resources

- ▶ Norsk Ordbank
 - ▶ full form dictionaries for Nynorsk and Bokmål; 106,789 and 142,899 lemmas, respectively
- ▶ The Oslo–Bergen tagger
 - ▶ Constraint Grammar morphological disambiguation
 - ▶ Constraint Grammar syntactic dependency parser
 - ▶ Various other modules (compounding, NER, ...)
- ▶ No freely available bilingual dictionary between Nynorsk and Bokmål, until now...

Reuse of Free Resources in Nynorsk↔Bokmål MT

Kevin Unhammer, Trond Trosterud

Introduction

Nynorsk and Bokmål

Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG conversion

Translation dictionary
Structural transfer

Evaluation

Coverage

WER and BLEU

Future work

The apertium-nn-nb pipeline

- ▶ Morphological analysis
 - ▶ Ittoolbox: XML format, compiles to very fast FSTs
 - ▶ one XML dictionary gives both analysis and generation
- ▶ CG pre-disambiguation
- ▶ Statistical disambiguation (HMM)
- ▶ Bilingual dictionary for lexical transfer
- ▶ Shallow syntactic transfer rules
 - ▶ Local re-ordering (det noun → noun det)
 - ▶ Insertions, deletions and substitutions of lexical units (and chunks, but we don't use them yet)
- ▶ Morphological generation (again with Ittoolbox)

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion

Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Constraint Grammar

- ▶ Rules work on ambiguous input and may SELECT one analysis over all others, or REMOVE one analysis from the set of analyses, or ADD a new tag, etc.
- ▶ Often thousands of short, hand-written rules
- ▶ Rules apply based on “context conditions”:
 - ▶ (-1^* noun) means “there must be word with a noun analysis somewhere to the left”
 - ▶ $(1C^* \text{ verb})$ means “there must be a word *disambiguated* to a verb somewhere to the right”
 - ▶ $(1^* \text{ verb LINK } 2 \text{ noun})$ means “there must be a verb-analysis to the right, and a noun-analysis two positions to the right of that”
 - ▶ $(1^* \text{ verb BARRIER noun})$ means “there must be a verb-analysis to the right, and no noun-analyses before that”
 - ▶ There are many other possibilities. . .

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Example of a CG rule

If input contains the word 'walks' analysed as either verb 3sg present or noun pl, the following rule

```
SELECT (verb 3sg present) IF
    (-1*C 3sg BARRIER verb)
    (NOT -1 det);
```

would choose the verb analysis if there is a disambiguated word, analysed as third singular, to the left, with no verb between the two; *and* there is no determiner to the left

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Development of apertium-nn-nb

- ▶ Most of the work done within 12 weeks (Google Summer of Code 2009)
- ▶ Helped by high quality free resources
 - ▶ Monolingual dictionaries: Norsk Ordbank converted from full form listing to Ittoolbox format
 - ▶ CG: Oslo–Bergen tagger converted to use Apertium tag scheme

Reuse of Free
Resources in
Nynorsk ↔ Bokmål
MT

Kevin Unhammer,
Trond Trosterud

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Disambiguation and CG conversion

Reuse of Free
Resources in
Nynorsk↔Bokmål
MT

Kevin Unhammer,
Trond Trosterud

- ▶ Bigram HMM's trained on Wikipedia text (Baum-Welch, 8 iterations)
- ▶ Conversion of CG tag set mostly done within a few days
- ▶ Errors fixed in CG reported back to Oslo–Bergen tagger team, win-win.
- ▶ However: the Oslo–Bergen tagger was designed for corpus annotation and lexicography
 - ▶ For the linguist, recall is more important than precision
 - ▶ For (our) MT, only one analysis matters
 - ▶ So we need to take more chances with our rules
 - ▶ Also, we get some MT-specific rules (like CG-based lexical selection)

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion

Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Finding word translations semi-automatically

- ▶ Method 1: *Exact matches where the morphology is the same*
 - ▶ If lemma and morphological possibilities are the same, assume we have a translation
 - ▶ ‘snøvla’, verb, pres/pass/imp/pret/inf... exists in both monolingual dictionaries; add it as a translation
 - ▶ 36,000 entries (although quite a lot are low-frequency / loan-words)
 - ▶ Risk of “radical forms”

Reuse of Free Resources in Nynorsk ↔ Bokmål MT

Kevin Unhammer, Trond Trosterud

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG conversion

Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Finding word translations semi-automatically

- ▶ Method 1: *Exact matches where the morphology is the same*
 - ▶ If lemma and morphological possibilities are the same, assume we have a translation
 - ▶ 'snøvle', verb, pres/pass/imp/pret/inf... exists in both monolingual dictionaries; add it as a translation
 - ▶ 36,000 entries (although quite a lot are low-frequency / loan-words)
 - ▶ Risk of "radical forms"
- ▶ Method 2: *Predictable substring-translations*
 - ▶ find Bokmål entries without translations
 - ▶ run string replacements for typical differences (-hjem→-heim-, -lig→-leg, ...)
 - ▶ check if the altered entries are in the Nynorsk analyser
 - ▶ ... and vice versa
 - ▶ Main run gave 2500 good entries

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG conversion

Translation dictionary

Structural transfer

Evaluation

Coverage

WER and BLEU

Future work

Expanding the translational dictionary using alignments

- ▶ Method 3: *Automatic word alignments*
 - ▶ Corpora:
 - ▶ KDE4 software translations (400,000 words)
 - ▶ government web pages (50,000 words, crawled with bitextor)
 - ▶ po-terminology (only on KDE4)
 - ▶ gave some hundreds of new terms
 - ▶ morphological tagging → Giza++ → ReTraTos
 - ▶ about 3500 entries
 - ▶ Lots of cleaning needed

Reuse of Free
Resources in
Nynorsk ↔ Bokmål
MT

Kevin Unhammer,
Trond Trosterud

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion

Translation dictionary

Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Expanding the translational dictionary using alignments

- ▶ Method 3: *Automatic word alignments*
 - ▶ Corpora:
 - ▶ KDE4 software translations (400,000 words)
 - ▶ government web pages (50,000 words, crawled with bitextor)
 - ▶ po-terminology (only on KDE4)
 - ▶ gave some hundreds of new terms
 - ▶ morphological tagging → Giza++ → ReTraTos
 - ▶ about 3500 entries
 - ▶ Lots of cleaning needed
- ▶ Method 4: User-contributed entries (via Wikipedia)

► Genitive noun phrases

- (2) a. forfatterens siste utgivelse
author.DEF.GEN last publication.IND
- b. den siste utgjevinga til forfattaren
the last publication.DEF of author.DEF
'the author's last publication'
- c. mitt nye luftputefartøy
my new hovercraft.IND
- d. det nye luftputefartøyet mitt
the new hovercraft.DEF mine
'my new hovercraft'

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion

Translation dictionary

Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

Evaluation

- ▶ Coverage
- ▶ WER
- ▶ BLEU

Reuse of Free
Resources in
Nynorsk↔Bokmål
MT

Kevin Unhammer,
Trond Trosterud

Introduction

Nynorsk and Bokmål

Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion

Translation dictionary
Structural transfer

Evaluation

Coverage

WER and BLEU

Future work

Coverage

- ▶ Naïve coverage on Nynorsk Wikipedia: 89.6%
- ▶ Naïve coverage on Bokmål Wikipedia: 88.2%
- ▶ Coverage seems to be the most important issue:
Not only is every 10th word untranslated, but we get disambiguation problems and transfer problems in the rest of the sentence

Reuse of Free
Resources in
Nynorsk↔Bokmål
MT

Kevin Unhammer,
Trond Trosterud

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

WER and BLEU scores in the nb→nn direction

- ▶ Word Error Rate, BLEU and Unknown Word Rate on text from government web pages

	BLEU	WER _O	WER _W	UWR
Apertium	0.74	32.5 (36.1)	17.7 (50.5)	9.5
Nyno	0.85	29.1 (34.6)	13.3 (47.3)	0.8

Table: BLEU score (two reference translations) and WER (for the Original and Wikipedia references). Numbers in parenthesis give percentage of unknown words which were free-rides.

- ▶ WER on post-edited Apertium MT output on a Wikipedia article, however, was 10.71% (64.93% free-rides)
- ▶ Coverage seems like the major difference.

Future work

► Compounding

- (3) a. bilkirkegård → bilkyrkjegard
car.cemetery → car.cemetery
- b. postordrelager → #postordrelagar
mail.order.storage → mail.order.creator

Reuse of Free
Resources in
Nynorsk ↔ Bokmål
MT

Kevin Unhammer,
Trond Trosterud

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

► Compounding

- (3) a. bilkirkegård → bilkyrkjegard
car.cemetery → car.cemetery
- b. postordrelager → #postordrelagar
mail.order.storage → mail.order.creator

► Multi-word expressions

- (4) a. Han anbefalte meg å gå hjem
he recommended me INF go home
- b. Han rådde meg til å gå heim
he counseled me to INF go home
'He recommended that I go home'

► Compounding

- (3) a. bilkirkegård → bilkyrkjegard
car.cemetery → car.cemetery
- b. postordrelager → #postordrelagar
mail.order.storage → mail.order.creator

► Multi-word expressions

- (4) a. Han anbefalte meg å gå hjem
he recommended me INF go home
- b. Han rådde meg til å gå heim
he counseled me to INF go home
'He recommended that I go home'

► Expanding the Scandinavian language group

Thanks for listening!

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work

This presentation may be distributed under the terms of the GNU GPL, GNU FDL and CC-BY-SA licences.

- ▶ GNU GPL v. 3.0
<http://www.gnu.org/licenses/gpl.html>
- ▶ GNU FDL v. 1.2
<http://www.gnu.org/licenses/gfdl.html>
- ▶ CC-BY-SA v. 3.0
<http://creativecommons.org/licenses/by-sa/3.0/>

Introduction

Nynorsk and Bokmål
Norwegian language resources

The Apertium architecture and nn-nb pipeline

Constraint Grammar

Developing apertium-nn-nb

Disambiguation and CG
conversion
Translation dictionary
Structural transfer

Evaluation

Coverage
WER and BLEU

Future work