Exploiting Parallel Texts to Leverage the Manual Annotation Bottleneck

The MultiSemCor case

Emanuele Pianta (ITC–irst, Trento)

Multi-lingual semantic annotation: Theory and applications,
Saarland University, Saarbruecken, Germany, June 26th and 27th 2006

Outline

- The manual annotation bottleneck
- Translation as annotation
- Cross-language annotation transfer (MultiSemCor)
- Critical issues and Evaluation
- Dictionary-based word alignment (Knowa)

The manual annotation bottleneck

- *Manual-quality* annotated resources are crucial for data-driven approaches to NLP
- Manual annotation is VERY costly
  - annotation schemes
  - annotators
  - time
- Alternatives
  - Use less annotated data (weakly supervised learning, self- and co-training …)
  - Reduce the cost of manual annotation

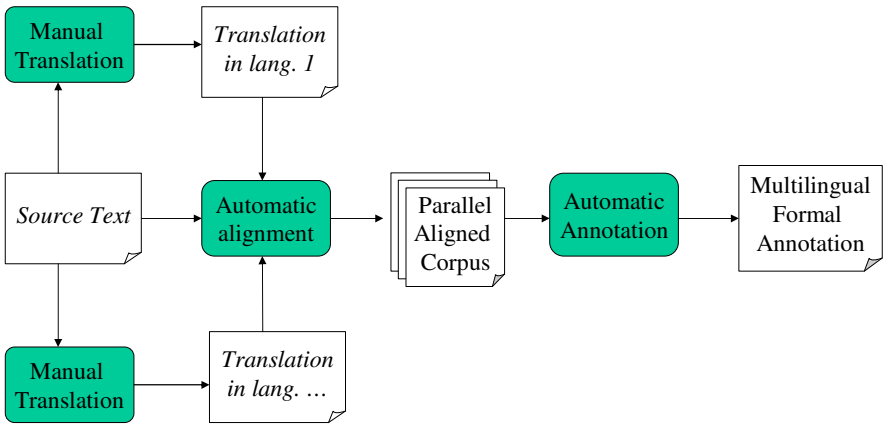
Annotation vs Translation

- *Annotation*: expressing some linguistic aspect of a text in a formal language
- *Translation*: expressing the content of a text in another natural language
- Is translation a sort of annotation?
  - Yes: mapping between two languages
  - No: formal vs informal
  - No: analytical vs synthetic

Translation as annotion (what if)

- Plenty of existing translations (parallel corpora)
- Plenty of translators
  - No need for special training
  - Less time
  - More rewarding and less error prone than formal annotation

The *translation as annotation* scheme



Background

- Word sense disambiguation
  - Brown et al. 1991, Gale at al. 1992
- Target word selection
  - Dagan et al. 1991, Dagan and Itai 1994
- Word sense clustering
  - Ide et al. 2002
- Cross-language word sense annotation
  - Diab 2002
- Tag projection + processor induction
  - Yarowsky et al. 2001
- Projection of syntactic relations
  - Hwa et al. 2002, Cabezas et al. 2001

Cross-Language Semantic Annotation Transfer

Create a semantically annotated corpus by exploiting information contained in an already annotated corpus, using word alignment as a bridge

Applied in the creation of the MultiSemCor corpus

Semantic annotation transfer

Observation:

- Translations are supposed to be *meaning preserving*

Hypothesis:

- if a source text has been *semantically annotated* and *aligned* to its translation
- then it is possible to *transfer* the annotation from the source text to its translation using word alignment as a bridge

MultiSemCor: the procedure

- Take the SemCor corpus
  - English, part of the Brown Corpus
  - ~700,000 running words, PoS annotated
  - 200,000 content words manually annotated with WordNet senses (and lemmas)
- *Step 1*: Manually *translate* SemCor into Italian
- *Step 2*: Automatically *align* the texts at the sentence and word level
- *Step 3*: Automatically *transfer* the word sense annotations from English to the aligned Italian words

MultiSemCor: *the result*

- an *Italian corpus* annotated with:
  - PoS
  - lemma
  - word sense
- an *English/Italian parallel corpus*
  - *aligned* at the word level
  - *lexically annotated* with a shared inventory of word senses (MultiWordNet synsets)

SemCor Corpus details

- 352 English texts
  - 676,546 running words, PoS annotated
  - 234,136 tagged with lemmas and WordNet senses
- *All PoS* tagged: 186 texts
  - 359,732 tokens – 192,639 semantically tagged
- *Only verbs* tagged: 166 text
  - 316,814 tokens – 41,497 semantically tagged

Step 1: Getting Translations

- In MultiSemCor Italian translations were made on purpose by professional translators

  Advantages:

  – translating and transferring annotation may be preferable to hand-labeling from scratch

  – a parallel corpus is produced

  – translations can be "controlled", i.e. carried out following criteria aiming at maximizing alignment and annotation transfer

Step 1: Controlled transaltion criteria

- Maintain *sentence segmentation*
- Mark Italian *multiword named entities* with an underscore
  – European_Union -> Unione_Europea
- Prefer the *Collins dictionary* (used by KNOWA)
- Choose the most "*synonymous" translation* equivalent (prefer same PoS)
- *Never to the detriment of a good Italian prose*

Step 2: Word Alignment

- KNOWA (KNOwledge-intensive Word Aligner)

  relies on:
  - English/Italian bilingual dictionary (Collins)
  - Morphological analyzer (TextPro)
  - Multiwords recognizer

- KNOWA for MultiSemCor
  - in SemCor all multiwords included in WordNet are explicitly marked
  - only content words have a semantic annotation

Step 3: Automatic sense transfer

- Given two aligned words, project the word sense annotation from the source to the target word.
- Requires a mapping between English and Italian synsets
  - Could be EuroWordNet Inter Lingual Index.
  - In MultiSemCor English and Italian *correspondent synsets* have the same identifier
- Relies on the hypothesis that translation preserves word meaning

Quality issues

Possible degradation of the quality of the resulting Italian annotation

1. *SemCor quality*:
   - annotation errors in the original corpus
2. *Word alignment quality*:
   - word aligner performance
3. *Transfer quality*: non-transferable annotations
   - non synonymous translation equivalents
     meaning -> motivo (reason)
   - synonymy at phrasal level but not at lexical level
     a dreamer sees -> una persona sogna (a person dreams)
   - translation equivalents which are not lexical units
     successfully -> con successo (with success)

MultiSemCor Gold Standard

- *4 English unseen texts* from SemCor
  - For each English text both a *free* and a *controlled* translation
  - 8,877 English tokens
  - 9,224 Italian tokens (controlled)
  - *[A]* Manual word *alignment*
  - *[B]* Manual Italian *word sense annotation*

[A] Kinds of word level alignments

	English	Italian
Simple words	health	salute
Segments (multiwords)	rain dance	danza della pioggia
Segments (generic phrases)	open-mouthed	con la bocca spalancata
Parts of words	clasping *him*	afferrando*lo*

[A] Kinds of semantic correspondences between aligned units

	English	Italian
Full (synonymic)	science	scienza
Non-synonymic	meaning	motivo (*reason, grounds*)
Trans-PoS non synonymic	dream previsions	sogni premonitori (*premonitory dreams*)
Fuzzy	the dreamer sees	una persona sogna (*a person dreams*)
Involving extra-grammatical elements	my hands	(le) mie mani
	(he) wants	vuole
omissions	the (ocean of) mankind	il genere umano

[B] Italian word sense annotation

- Look at the *annotation* of the English *aligned* word
- If the English synset is appropriate for the Italian word, then transfer the annotation
- Otherwise look for the right synset in MultiWordNet
- Distinguish
  - *errors* in the SemCor annotation
  - *non transferable* annotations

Gold Standard

	ENGLISH				ITALIAN		ALIGN
English word	*English segment*	*SemCor synset*	*Italian PoS*	*Italian word*	*Italian segment*	*Gold Standard Synset*	*Word alignment*
1-There	1:mw		cli	1-Ci	1:mw	1=v#01775163	1-There
2-are	1:mw	v#01775163	v	2-sono	1:mw		2-are
3-certainly		r#00140819	p	3-,	p:nt		
4-large		a#01328712	adv	4-certamente		r#00140819	3-certainly
5-areas		n#04596100	p	5-,	p:nt		
6-of			adj	6-vasti		a#01328712	4-large
7-understanding	2	*n#04473075*	n	7-ambiti		n#04596100	5-areas
8-in			prep+art	8-della	2:gr:art=>la		6-of=>di
9-the			n	9-conoscenza	2	*n#00013243*	7-understanding
10-human	3:mw	a#02575805	prep+art	10-nelle			[8-in=>in, 9-the=>le]
11-sciences	3:mw	n#04596907	n	11-scienze	3:mw	n#04596907	11-sciences
12-which			adj	12-umane	3:mw	a#02575805	10-human
…	…	…	…	…	…	…	…

Inter-annotator agreement

- Agreement : $\frac{\text{2 N of common words}}{\text{N of aligned/annotated words}}$

- *[A]* Agreement on Word Alignment (Free): 87%
- *[A]* Agreement on Word Alignment (Controlled): 92%
- *[B]* Agreement on Sense Annotation: 81.9%
  (cfr. SemCor Agreement: 78.6%)

We can know assess Quality Issues

1. *SemCor quality*:
   - annotation errors in the original corpu
2. *Word alignment quality*:
   - word aligner performance
3. *Transfer quality*:
   - non-transferable annotations

1. SemCor errors

"He put his hands on his *pockets*"

Annoted as: {pouch, sac, sack, pocket -- *an enclosed space*}

Right: {pocket--*a small pouch in a garment for carrying small articles*}

SemCor errors estimated on the MultiSemCor GS: 2.8%

2. KNOWA Evaluation

- Precision $\frac{\text{Knowa correct alignments}}{\text{Knowa alignments}}$
- Recall $\frac{\text{Knowa correct alignments}}{\text{Gold Standard alignments}}$
- Coverage $\frac{\text{Knowa alignments}}{\text{Gold Standard alignments}}$

Knowa vs GIZA++ on the EuroCor Task

- Training set: 22,000 aligned sentences

		Precision	Recall	F-measure
GIZA++ Ita -> Eng	all	62.0	34.7	44.5
	content	53.2	38.3	44.5
KNOWA basic	all	63.4	41.6	50.2
	content	85.5	53.2	65.5

Knowa vs GIZA++ on the MultiSemCor Task – Development Set

	task	Prec.	Recall	F-meas
GIZA++ 25k (MSC+EC) Italian ->Eng	all	73.4	55.2	63.0
	semcor	81.9	52.9	64.3
GIZA++ 25k (MSC+EC) Intersection	all	95.2	38.8	55.1
	semcor	95.8	37.1	53.5
KNOWA basic	all	84.5	63.7	72.6
	semcor	92.0	73.4	81.7

2. (Best) KNOWA on MultiSemCor

	Full Text			Annotated-only		
	P	R	C	P	R	C
Free	85.9	61.8	70.0	92.8	70.3	75.8
Controlled	89.2	69.4	76.1	***94.7***	***76.2***	***80.5***

2. KNOWA Evaluation (PoS break-down)

Translation	PoS	Precision (%)	Recall (%)	Coverage (%)
Controlled	Nouns	96.9	84.5	87.2
	Verbs	91.4	77.4	84.7
	Adjectives	96.0	72.3	75.3
	Adverbs	91.0	54.4	59.8
	Total	94.7	76.2	80.5

3. Non-transferable annotations

- The translation equivalent is not synonymous
- The translation equivalent is not a lexical unit

3.1 Non synonymous translations

- Translation equivalents which are not cross-language synonyms:
  - "meaning" → "motivo" (lit. reason, grounds)
- Change of lexical category:
  - "the possibility for man to *coexist* with animals"
    → "le possibilità di *coesistenza* tra gli uomini e gli animali"
- Translation errors:
  - "pull" → "spingere" (lit. push)
- Phrase level correspondence:
  - "a dreamer sees" → "una persona sogna"

3.2 Non lexical-unit translations

- Lexical gaps:
  - "successfully" → "con successo"
- Translator choice
  - "empirically" →
    "in modo empirico" vs "empiricamente"

How many non-transferable annotations?

- 16.9% of the English annotations
  - 85.4% non synonyms
  - 14.6% non lexical units

Final Italian Annotation Quality

- Total Italian words to be annotated: 4,313
- Transferred senses: 3,297
  - correct: 2,897
  - incorrect: 400
- Remaining Italian words to be annotated: 1,016

	Precision	Recall	Coverage
Controlled	***87.9***	67.2	76.4

Error Rate: 12.1

Error analysis (12.1%)

	Theoretical Error	Actual Error #	Actual Error %
SemCor errors	2.8% (117)	109	3.3
Word alignment errors	5.3%	95	*2.9
Non-transferable annot. (non-synonymous only)	14.4%	196	5.9
Total Incorrect transfer		400	12.1

* only on transferable annotations

The MultiSemCor Corpus so far

- *116* English texts aligned at the word level with their corresponding Italian translations
  - 251, 414 English tokens
  - 261,887 Italian tokens
  - 116,826 English semantic tags
  - 90,588 Italian semantic tags
- Freely distributed for research purposes
  in XML-based standard compliant format

The usage of MultiSemCor

- To automatically enrich the Italian component of MultiWordNet
  - out of 22,988 Italian word senses in MultiSemCor 15,943 are not present in MultiWordNet
- Gold standard for the evaluation of WSD systems working on Italian
- To be consulted by humans through a web interface available at: *http://multisemcor.itc.it*

Conclusion

Cross-language annotation transfer methodology

- An approach to the creation of high quality semantically annotated corpora based on the exploitation of parallel texts
  - exploits existing (mostly English) annotated resources
  - creates corpora in new (resource-poor) languages
  - reduced human effort

Future work

- Apply the methodology to the remaining 70 SemCor texts
- Enlarge the evaluation gold standard
- Extend the methodology to other languages, e.g. Spanish and Romanian, for which a WordNet exists and can be aligned with MultiWordNet
- Explore the possibility of transferring syntactic annotation
  - The Brown Corpus (of which SemCor is part) has been annotated within the Penn Treebank

The MultiSemCor Web Interface

- Intended for:
  - Lexicography
  - Translation studies
  - Linguistic teaching
  - Multilingual browsing
- Showing:
  - Linguistic annotation
  - Bilingual sentence alignment
  - Bilingual semantic concordancing
  - Integration between corpora and lexical resources (WordNet)

Two browsing modalities

- Text oriented (sentence alignment)
- Word oriented (semantic concordancer)

http://multisemcor.itc.it