ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT

Liane Guillou*, Christian Hardmeier†, Aaron Smith†, Jörg Tiedemann† and Bonnie Webber*

University of Edinburgh*, University of Uppsala†

L.K.Guillou@sms.ed.ac.uk, christian.hardmeier@lingfil.uu.se, aaron.smith.4159@student.uu.se, jorg.tiedemann@lingfil.uu.se, bonnie@inf.ed.ac.uk

Download: http://opus.lingfil.uu.se/ParCor

Problem: Pronominal Coreference in SMT

Pronominal Coreference: Pronouns are used in place of full referring expressions, e.g.

> I have an umbrella. **It** is red.
> "**It**" is the *referent* and "umbrella" is the *antecedent*
>
> I need an umbrella. **It** is raining.
> "**It**" does not refer to anything

Challenges for Statistical Machine Translation (SMT):

- Pronoun-antecedent agreement (e.g. in number and gender) for some languages
  > I have an umbrella. **It** is red.
  > Ich habe einen Regenschirm. **Es** ist rot. (Should be **Er** – Regenschirm {masc.})
- Pronoun may be used/required in one language but not the other (insertion/deletion)
  > **their** expertise, integrity, drive and hunger...
  > **ihrem** Fachwissen, **ihrer** Integrität, **ihrem** Bestreben und **ihrer** Bereitschaft...

What is the source of these problems? An annotated parallel corpus could help

Data: English Texts and German Translations

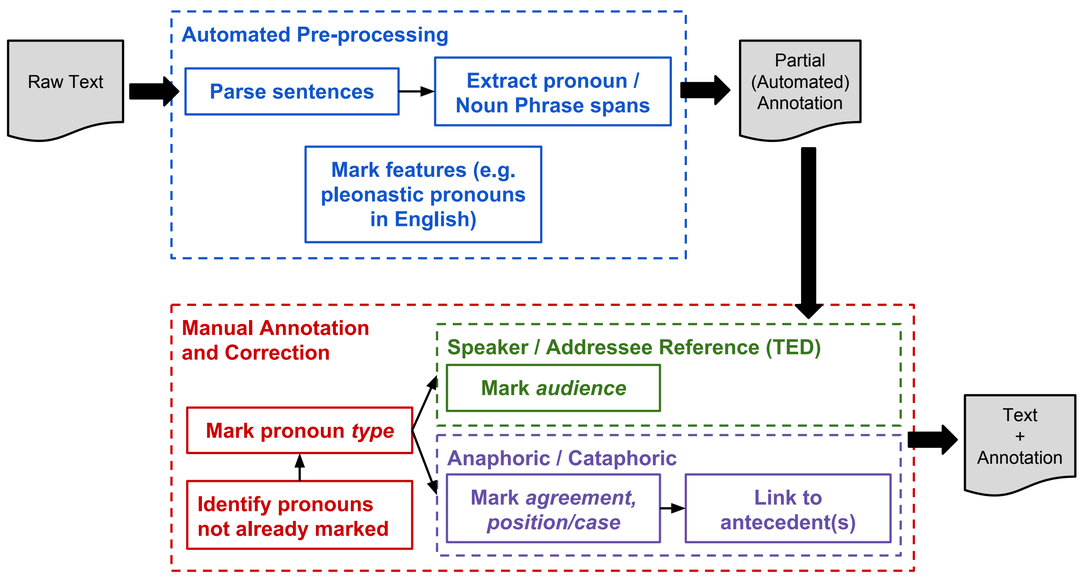
TED

- Orally delivered public lectures in front of a live audience
- 11 TED Talk transcriptions. Translations provided by volunteers
- Source: IWSLT 2013 MT shared task test sets (WIT^3 corpus [Cettolo et al., 2012])

EU Bookshop

- EU publications written for an educated but non-expert public
- 8 documents. Translations provided by professional translators
- Source: EU Bookshop online archive

Annotation Process



- Automated Pre-processing – reduce human annotation effort and improve Inter-Annotator Agreement (IAA)
  - Parser, pleonastic "it" detector, string match, rules
- Manual Annotation – start with partial annotations, annotated in MMAX-2 [Müller and Strube, 2006], according to guidelines based on MUC-7 [Chinchor and Hirschman, 1998]

Pronoun Types

- **Anaphoric reference** – refers to an antecedent noun phrase
- **Event reference** – refers to propositions, facts, states, situations, opinions, etc.
- **Pleonastic** – does not refer to anything (e.g. It is raining / Es regnet)
- **Addressee reference** – refers to the person(s) being addressed (e.g. "you")
- **Speaker reference** – refers to the speaker / writer (e.g. "I", "we")
- **Other function** – includes cataphoric, generic and extra-textual reference pronouns and those for which the type could not be determined

Pronoun Features

Anaphoric/cataphoric Reference:

- **Agreement** – number, gender and politeness (ambiguous pronouns only)
- **Position** – subject / non-subject (English only). Grammatical **Case** (German only)
- Link to nearest non-pronominal antecedent(s)

Addressee/speaker Reference:

- **Audience** – whether "we/you" includes the audience (TED Talks only)

Corpus Statistics

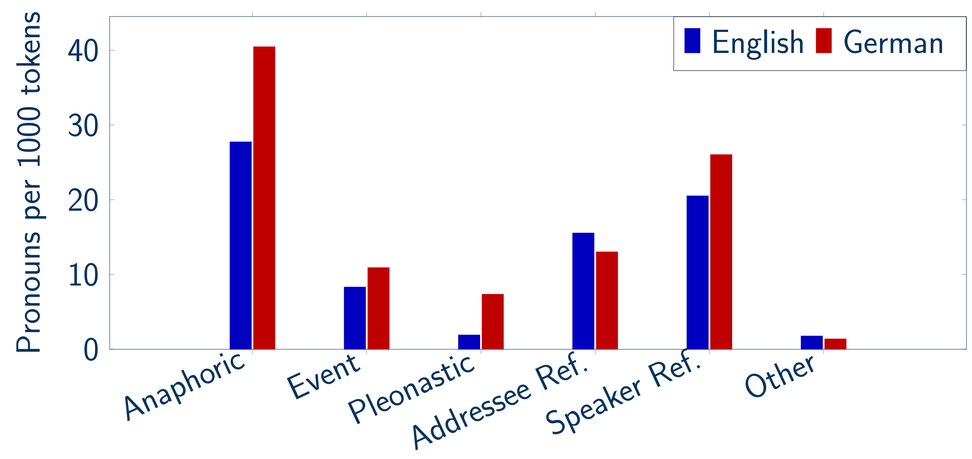


Figure : Pronoun **type** counts for English and German **TED** Talks

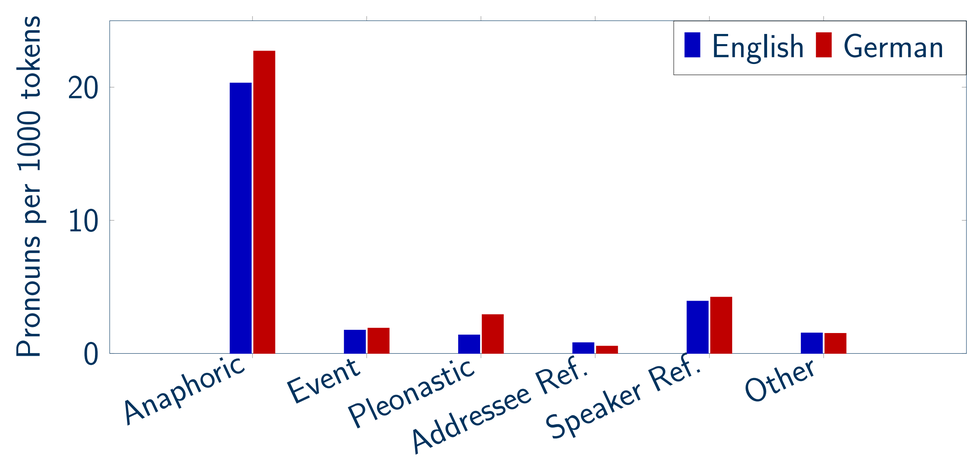


Figure : Pronoun **type** counts for English and German EU **Bookshop** documents

Inter-Annotator Agreement

- Measured using Cohen's Kappa [Cohen, 1960] for: Pronoun *type*, agreement, position (English only), case (German only) and audience (TED Talks only)
- Scores computed for pronouns annotated by 2 annotators
- Antecedents are spans. IAA considers exact and partial matches

Category	Pronouns	Disagree	Kappa
ENGLISH:			
Type	138	13	0.85
Agreement	73	0	1.00
Position	73	5	0.82
Antecedent	73	13	N/A
GERMAN:			
Type	205	4	0.96
Agreement	136	4	0.96
Case	136	11	0.85
Antecedent	136	9	N/A

Table : IAA Scores for EU Bookshop document "MJ3011331"

Category	Pronouns	Disagree	Kappa
Type	363	37	0.85
Agreement	133	6	0.90
Position	133	2	0.98
Antecedent	133	10	N/A
Audience	163	22	0.75

Table : IAA Scores for English TED Talk with ID "824"

German TED annotation was provided by a single annotator

Future Work

- Use the corpus to build SMT systems with a specific focus on improving the translation of pronoun coreference
- Continue working on corpus development:
  - Additional TED Talks / EU Bookshop publications, new languages and text genres
  - Expand the capabilities of the automated pre-processing pipelines

Acknowledgements

Manual annotations were provided by Susanne Tauber, Petra Strom, Samuel Gibbon and David Lawrence and the German pre-processing pipeline was provided by Yannick Versley. The work carried out at Edinburgh University was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE). The work at Uppsala University was supported by the Swedish Research Council (Vetenskapsrådet) through the project on Discourse-Oriented Machine Translation (2012-916).

Bibliography

M. Cettolo, C. Girardi and M. Federico (2012). WIT^3: Web Inventory of Transcribed and Translated Talks. *Proceedings of EAMT 2012, pages 261–268.*

N. Chinchor and L. Hirschman (1998). MUC-7 Coreference Task Definition (v3.0). *Proceedings of MUC-7.*

J. Cohen (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20(1): pages 37–46.*

C. Müller and M. Strube (2006). Multi-Level Annotation of Linguistic Data with MMAX2. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pages 197–214. Peter Lang, Germany.*