

Pronoun Translation

Liane Guillou and Alexander Fraser
{liane,fraser}@cis.uni-muenchen.de

CIS, Ludwig-Maximilians-Universität München

Morphology

2016

13.06.2016

Outline

1. Introduction
2. Machine Translation
3. Pronoun Translation in Statistical Machine Translation
4. Cross-lingual Pronoun Prediction
5. Other Issues

1. Introduction
2. Machine Translation
3. Pronoun Translation in Statistical Machine Translation
4. Cross-lingual Pronoun Prediction
5. Other Issues

Introduction

Pronouns

- Examples: I, you, he/she, it, they, this, that
- First classified as a part of speech in 2BC (Dionysius Thrax, Hellenistic grammarian)
- Early definition: pronouns are a “noun substitute”
- Universal to language (Greenberg 1963)
- Pronouns occur at the *discourse level*
 - **Discourse**: coherent sequences of sentences, propositions, speech, or turns-at-talk
- Different ways to categorise pronouns: form vs. function

Introduction

Pronouns by Form

Commonly seen in grammar books:

- **Personal:** Classified by person [1st/2nd/3rd], number [sg./pl.], case
- **Possessive:** Indicates possession, e.g. “Hey! *That’s mine!*”
- **Reflexive:** e.g. “John talks to *himself*”
- **Reciprocal:** e.g. “The boys don’t like *each other*”
- **Demonstrative:** e.g. “Could you pass me *that?*”
- **Indefinite:** Refers to unspecified persons/things, e.g. “*Everyone* likes cats”
- **Relative:** Used in relative clauses, e.g. “That’s the lady *who* owns the pie shop”
- **Interrogative:** Used to ask questions, e.g. “*Who* said that?”

Introduction

Pronouns by Function

- Alternative: Categorise pronouns by the function that they perform
- I.e. a “Functional Grammar” approach
- Some pronoun functions:
 - **Speaker**: Refers to the speaker, e.g. “*I* like cats”
 - **Addressee**: Refers to the reader/audience, e.g. “How are *you*?”
 - **Generic**: Refers to people in general, e.g. “If *you* own a car, *you* must insure it”
 - **Pleonastic**: Used to fill the subject position slot, e.g. “*It* is raining”
 - **Extra-textual**: Refers to things not in the text, e.g. “Look at *that*!”
 - **Event-reference**: Refers to a verb, verb phrase, clause, sentence

Introduction

The Anaphoric Function

- Anaphoric pronouns *corefer* with a noun phrase

Example

I have an [umbrella]₁. [It]₁ is red.

- **Anaphoric pronoun:** “it”
- **Noun phrase:** “umbrella” (also called the *antecedent*)
- **Corefer:** “referring to the same thing”

Introduction

Translating Anaphoric Pronouns

- In languages with grammatical gender, pronoun and antecedent must agree in **number** and **gender**
- **Number**: singular, plural
- **Grammatical gender**:
 - German [3]: masculine, feminine, neuter
 - French [2]: masculine, feminine
 - Czech [4]: masculine animate, masculine inanimate, feminine, neuter
 - English: none
- Grammatical gender contrasts with **natural gender**: male, female
 - In English we have the pronouns “he” and “she”

Introduction

Translating Anaphoric Pronouns

- In languages with grammatical gender, pronoun and antecedent must **agree in number and gender**

German example: umbrella → Regenschirm [masc. sg.]

I have an umbrella. [It] is red.

Ich habe einen Regenschirm. [Es] ist rot.

Ich habe einen Regenschirm. [Sie] ist rot.

Ich habe einen Regenschirm. [Er] ist rot. ✓

Introduction

Translating Anaphoric Pronouns

- Pronoun-antecedent agreement also required in: French, Spanish, Czech, Italian, etc.

French example: bicycle → vélo [masc. sg.]

I have a **bicycle**. [**It**] is red.

J'ai un **vélo**. **Il** est rouge. ✓

J'ai un **vélo**. **Elle** est rouge.

- Other antecedent translations are possible, but agreement must hold

French example: bicycle → bicyclette [fem. sg.]

I have a **bicycle**. [**It**] is red.

J'ai une **bicyclette**. **Il** est rouge.

J'ai une **bicyclette**. **Elle** est rouge. ✓

Introduction

Functional Ambiguity: “it” can perform many functions

Anaphoric: pronoun corefers with noun phrase

I have an umbrella. **[It]** is red.

Ich habe einen Regenschirm. **[Er]** ist rot.

Pleonastic: “dummy” pronoun fills subject position

I have an umbrella. **[It]** is raining.

Ich habe einen Regenschirm. **[Es]** regnet.

Event reference: pronoun refers to span of text containing a verb

X invaded Y. **[It]** resulted in war.

X besetzte Y. **[Dies]** führte zu Krieg.

Introduction

Position / Case

- **Position:** subject / object [English]
- **Case:** nominative, accusative, dative, etc.

Case	Singular			Plural
(English nominative)	he	she	it	they
Nominative (subject)	er	sie	es	sie
Accusative (direct object)	ihn	sie	es	sie
Dative (indirect object)	ihm	ihr	ihm	ihnen
Genitive	seiner	ihrer	seiner	ihrer

Table : Third-person German Pronouns

- Some languages have many cases:
 - Czech [7]: nom, acc, dat, gen, *vocative* (indicates person/thing being addressed), *locative* (indicates location), *instrumental* (noun is the means by which the subject accomplishes an action)

Introduction

Case

- English largely lost its case system but personal pronouns retain it
 - **Subject:** I, he, we (e.g. “I kicked the ball”)
 - **Object:** me, him, us (e.g. “He kicked *me*”)
- Case determined by the *grammatical function* or *syntactic role* that the pronoun performs
 - **Nom:** subject of a finite verb (“I went to the cinema”)
 - **Acc:** direct object of a verb (“The clerk remembered *me*”)
 - **Dat:** indirect object of a verb (“She gave a discount to *me*”)
 - **Gen:** indicates possession (“That book is *mine*”)

- In German, the preposition also determines the case of nouns, pronouns, adjectives
 - E.g. “mit” always takes the dative case
 - Two-way prepositions may take dative or accusative e.g. “in”
 - Die Leute gehen in die Kirche. [motion: acc]
 - Die Leute sitzen in der Kirche. [location: dat]

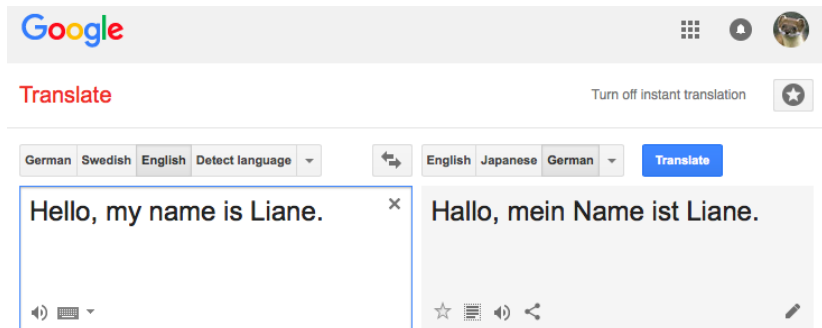
Outline

1. Introduction
2. Machine Translation
3. Pronoun Translation in Statistical Machine Translation
4. Cross-lingual Pronoun Prediction
5. Other Issues

Machine Translation

Introduction

- The use of software to translate text from one language to another
- E.g. Google Translate



The screenshot shows the Google Translate web interface. At the top left is the Google logo. To the right are icons for a grid, a notification bell, and a profile picture. Below the logo is the word "Translate" in red. To the right of "Translate" is the text "Turn off instant translation" and a star icon. The main interface has a language selection bar with "German", "Swedish", "English", and "Detect language" buttons. A swap button is in the center, and "English", "Japanese", and "German" are selected on the right. A blue "Translate" button is on the far right. The input text box on the left contains "Hello, my name is Liane." and has a close button (x) in the top right corner. Below the input box are icons for a speaker and a keyboard. The output text box on the right contains "Hallo, mein Name ist Liane." and has icons for a star, a list, a speaker, and a share icon. A pencil icon is in the bottom right corner of the output box.

Machine Translation

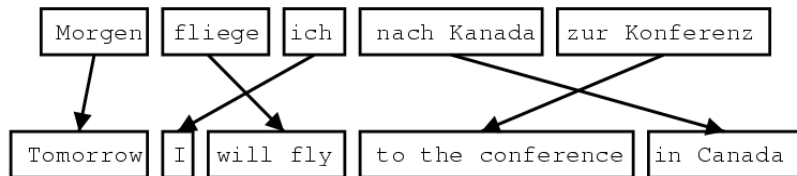
The Early Days

- **Word-based:** look up each word in a dictionary [50s]
 - Pro: Simple
 - Con: Words selected out of context
 - Con: Cannot handle idiomatic expressions (e.g. “Bite the dust”)
- **Rule-based:** use morphological and syntactic rules [70s]
 - Pro: Linguistically motivated → better translations
 - Con: Systems quickly become complex and difficult to maintain
- Example rule for English-to-French translation:
NP : Adjective Noun → NP : Noun Adjective
E.g. black cat → chat noir

- **Statistical MT**: use large amounts of parallel data to train systems
- *Statistical*: work out the probability of a word / words being translated as X based on frequencies in the parallel data
- Split documents into sentences, which are then translated in isolation
- Common paradigms:
 - **Phrase-based**: translate phrases (i.e. sub-strings)
 - **Syntax-based**: translation to/from syntax trees
- 2000s to present

Machine Translation

Phrase-based Models



- German input is segmented into **phrases**
 - Any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Translation is built from left-to-right
- Phrases may be reordered

Machine Translation

Phrase-based Models

- **Pro:** Rather simple
- **Pro:** Can handle idiomatic expressions
- **Pro:** Produced state-of-the-art translation quality for many language pairs (until recently)
- **???:** Method is not linguistically motivated
- **Con:** Reliant on lots of parallel data
 - Ok for resource *rich* pairs, e.g. English-German
 - Not good for resource *poor* pairs, e.g. Jerrais-Gaelic
- **Con:** Sentences translated in isolation
 - Information from previous sentences not available when translating the current sentences

Outline

1. Introduction
2. Machine Translation
3. Pronoun Translation in Statistical Machine Translation
4. Cross-lingual Pronoun Prediction
5. Other Issues

Statistical Machine Translation

Problems: Anaphoric pronouns

Example

I have an [umbrella]₁. [It]₁ is red.

- Work has focussed on Statistical Machine Translation (SMT)
- Information from previous sentences not available when translating the current sentences
- SMT systems rely on small context window around pronoun to select translation
- **Inter-sentential**: pronoun and antecedent in different sentences
- **Intra-sentential**: pronoun and antecedent in the same sentence
- Both cases are a problem for SMT

Statistical Machine Translation

Example Translation: Intersentential Anaphoric Pronoun

The screenshot shows the Google Translate web interface. At the top left is the Google logo. To the right are icons for a grid, a notification bell, and a profile picture. Below the logo is the word 'Translate' in red. To the right of 'Translate' is the text 'Turn off instant translation' and a star icon in a square. Below this is a horizontal bar with language selection buttons: 'English', 'German', 'Swedish', and 'Detect language' with a dropdown arrow. In the center is a bidirectional arrow icon. To the right are buttons for 'German', 'English', and 'Japanese' with a dropdown arrow, followed by a blue 'Translate' button. Below the language bar are two text boxes. The left box contains the English text 'I have an umbrella. It is red.' with a close 'x' icon and a speaker icon with a keyboard icon below it. The right box contains the German translation 'Ich habe einen Regenschirm. Es ist rot.', with 'Es' highlighted in blue. Below the German text are icons for a star, a list, a speaker, a share icon, and an edit icon.

Statistical Machine Translation

Problems: Disambiguating Pronoun Function

- Different pronouns required to translate different functions of “it”:
 - Anaphoric: er, sie, es
 - Pleonastic: es
 - Event reference: dies, das
- SMT systems rely on small context window around pronoun to select translation
- Context window may not be enough to disambiguate pronoun function

Statistical Machine Translation

Some Possible Solutions

- Use **external tools** to detect:
 - Pronoun antecedents (anaphora/coreference resolution)
 - Pleonastic “it”
 - Position / case (dependency parser)
- Add this information to SMT pipeline:
 - **Pre-processing**: encode information in SMT training data [beginning]
 - **Decoding**: add a component within the SMT system [middle]
 - **Post-processing**: fix errors in SMT output [end]

Statistical Machine Translation

Anaphora / Coreference Resolution

- **Anaphora resolution:** find pronoun's antecedent

Example: anaphora resolution

I have **an umbrella**. **It** is red.

- **Coreference resolution:** find chains of coreferring pronouns / noun phrases

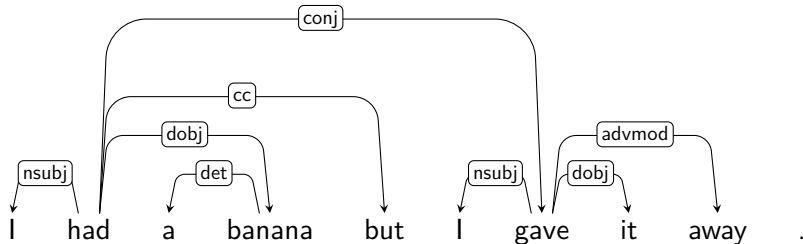
Example: coreference resolution

I have **an umbrella**. **The umbrella** is small and red. I use **it** when it rains.

Statistical Machine Translation

Dependency Parser

- **Dependencies:** words are connected to each other by directed links
- The (finite) verb is the structural centre or **root** ["had"]
- All other words are connected to the verb (directly/indirectly) by directed links



Statistical Machine Translation

Alternative Solutions

- Don't use discourse information:
 - Use larger context windows
 - Find other ways to span more text (sophisticated language models?)
- Use a rule-based MT system
 - Linguistic rules encode how to translate pronouns
 - May incorporate coreference resolution, pleonastic “it” detection, etc.

Statistical Machine Translation

DiscoMT 2015 Pronoun Translation Task (Hardmeier et al. 2015)

- **Shared Task:** teams compete on a common task
 - Build MT systems trained on common data
 - Translate a “test” file (“answers” are unknown)
 - Systems scored and ranked by shared task organisers
- DiscoMT 2015 shared task on pronoun translation
 - Translate subject position “it” and “they” into French
 - Score translations manually, and using automatic metrics (e.g. BLEU)

Statistical Machine Translation

DiscoMT 2015 Pronoun Translation: Systems

- Baseline: Basic phrase-based SMT system
- UU-Tiedemann: extension of the baseline, no discourse features
- IDIAP: classifier predicts pronoun translation
 - Use coreference resolution to identify anaphoric pronoun antecedents
 - Automatically replace pronouns in SMT system output (post-editing)
- UU-Hardmeier: classifier is an internal component of SMT system
 - Coreference resolution
- auto-postEDIt: rule-based automatic post-editing of SMT output
 - Coreference resolution
 - Focus on gendered anaphoric pronouns vs. non-anaphoric pronouns
- Its2: rule-based MT system
 - Coreference resolution
 - Focus on gendered anaphoric pronouns

Statistical Machine Translation

DiscoMT 2015 Pronoun Translation: Results

	Accuracy	BLEU
Official Baseline	0.676	37.18
IDIAP	0.657	36.42
UU-Tiedemann	0.643	36.92
UU-Hardmeier	0.581	32.58
auto-postEDIt	0.543	36.91
Its2	0.419	20.94
A3-108	0.081	4.06

Table : Official Shared Task Results

- **Accuracy**: pronouns match category: ce, ça/cela, il, ils, elles, elles or “other” (manual)
- **BLEU**: automatic measure of overlap between MT output and a human-authored reference translation (general-purpose)

Outline

1. Introduction
2. Machine Translation
3. Pronoun Translation in Statistical Machine Translation
4. Cross-lingual Pronoun Prediction
5. Other Issues

Cross-lingual Pronoun Prediction

Problem

- Break the translation problem down:
 - Build classifier to predict pronoun translation
 - (Later) incorporate classifier in MT system
- Cross-lingual pronoun prediction uses information from:
 - The human-authored source-language text
 - The target-language translation (human, MT)
- Aim: find the translation of each pronoun in the source-language text

Cross-lingual Pronoun Prediction

Example

Example: find French pronoun to replace XXX

They arrive first .

XXX arrivent en premier .

- What information do we have?
 - English pronoun: “they” [3rd-person pl., always subject position]
 - French verb: “arrivent” [3rd-person pl.]
- Translation is probably either “ils” [masc. pl.] or “elles” [fem. pl.]
- What information is missing?
 - Is “they” anaphoric, or generic?
 - If anaphoric, what is the antecedent? (for gender information)

Cross-lingual Pronoun Prediction

DiscoMT 2015 Pronoun Prediction Task (Hardmeier et al. 2015)

- Predict French translations of subject position “it” and “they”
- Nine prediction classes:
 - **ce**: primarily used as a “neuter” pronoun to refer to events/situations
 - **elle**: feminine singular subject pronoun
 - **elles**: feminine plural subject pronoun
 - **il**: masculine singular subject pronoun
 - **ils**: masculine plural subject pronoun
 - **ça**: demonstrative pronoun (“pick *that* up”)
 - **cela**: demonstrative pronoun (“that”)
 - **on**: indefinite pronoun (“*One* is most pleased...”)
 - **other**: some other word, or nothing at all, should be inserted

Cross-lingual Pronoun Prediction

Possible Features

- **Source-language text** [English]
 - Anaphora / coreference resolution
 - Pleonastic “it” detection
 - Dependency parse (subject / non-subject “it”)
 - X tokens either side of the pronoun
 - etc.
- **Target-language text** [French]
 - Morphological features of nearest verb
 - Morphological features of nearest noun
 - X tokens either side of the placeholder
 - etc.

Cross-lingual Pronoun Prediction

DiscoMT 2015 Pronoun Prediction: Training Data

Training data is supplied in a file format with five tab-separated columns:

- 1) The class label
- 2) The word actually removed from the text
- 3) The English source segment
- 4) The French **human authored** target segment with pronoun placeholders
- 5) List of source-target word alignments (numbers start at zero)

Example:

elles Elles They arrive first . REPLACE_0 arrivent en premier . 0-0 1-1 2-3 3-4

Cross-lingual Pronoun Prediction

DiscoMT 2015 Pronoun Prediction: Results

- Teams given a test file with predictions missing
- 13 systems + baseline (language model)
- Scores: macro-averaged F (across all prediction classes), accuracy

	Macro-F	Accuracy
Official Baseline	0.584	0.663
UU-Tiedemann	0.579	0.742
UEDIN	0.571	0.723
MALTA	0.565	0.740
...
IDIAP	0.164	0.407
A3-108	0.129	0.240
WITHDRAWN	0.122	0.325

Table : Official Shared Task Results

Cross-lingual Pronoun Prediction

WMT 2016: A New Shared Task

- Human authored target-language text is not realistic for MT setting
- **WMT 2016 task:** Replace words in human authored target-language text with POS tag + lemma

elles Elles They arrive first . REPLACE_0 arriver|VER en|PRP premier|NUM
.|. 0-0 1-1 2-3 3-4

- Simulates MT environment in which we can't trust morphological inflection
- Simulates two-step translation:
 - Step 1: **translate** English lemmas → French lemmas
 - Step 2: **generate** fully inflected French from French lemmas + features
- **Exercise:** WMT 2016 cross-lingual pronoun prediction task

Outline

1. Introduction
2. Machine Translation
3. Pronoun Translation in Statistical Machine Translation
4. Cross-lingual Pronoun Prediction
5. Other Issues

Other Issues

When to Translate Pronouns

- We cannot always assume that a pronoun in the source language should be translated as a pronoun in the target language
- It may be unnatural to use a pronoun
- It may be wrong to use a pronoun

Example: unnatural to use a possessive in German

Deshalb bleibt XyzTech mit positivem Cash Flow und gutem Ergebnis im Konzern.

As a result, we shall retain XyzTech, with **its** positive cash flow and good earnings.

Example from Becher (2011)

Other Issues

Other Translation Directions

- English → German/French: pick the correct gender or pronoun
- German → English: other problems exist
 - “sie” is ambiguous (“she”, “it”, “they”)
- French → English: different problems exist
 - “il” is ambiguous (“he”, “it”)
- Anaphora / coreference resolution can help again

Other Issues

Pro-drop

- **Pro-drop**: pronouns may be omitted when they are in some way inferable from the text
- Czech, Spanish and Japanese are pro-drop languages
- English, French and German are not
- BUT English subject pronouns often dropped in imperative sentences (e.g. Come here!)

Example: subject pro-drop in Czech

(1) I have an umbrella . **It** is red .
(Já) mám ∅ deštník . (On) je červený .
Mám deštník. Je červený.

- Czech to English: identify pro-drop in Czech, insert pronoun in English
- English to Czech: identify pronouns that should be dropped in Czech

Summary

- Some pronouns exhibit **functional ambiguity**: e.g. “it”
- Their correct translation requires *disambiguation* of function
- Problem of translating anaphoric pronouns into languages with *grammatical gender*
- Possible solutions to pronoun translation problem:
 - Incorporate external information (coreference resolution etc.)
 - Affecting different stages of the translation pipeline: pre, decoding, post
 - Cross-lingual pronoun prediction

Questions?

Thank you for your attention

- Viktor Becher (2011) Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts. PhD thesis, Department of Applied Linguistics (Institut für Sprachlehrforschung), University of Hamburg.
- Joseph H. Greenberg (1963) Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. Universals of Human language. MIT Press.
- Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley and Mauro Cettolo (2015) Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation. Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT).