Agreement matters: Challenges of translating into a MRL

Yoav Goldberg
Ben Gurion University
Why should we care about syntactic modeling of MRLs?
A brief summary
A brief summary

- What Kevin said.
Example: English ☞ Hebrew

I wash the car
Example: English ☐ Hebrew
I wash the car
אני רוחץ את
car
(I wash the car)
Example: English ➭ Hebrew

I wash the car
אני רוחץ את
(איני רוחץ את)

I wash the floor
Google translate
Google translate

(אאני רוחץ את)
(אאני רוחץ את)

(אאני רוחץ את)
(אאני רוחץ את)

(אאני רוחץ את)
(אאני רוחץ את)
Example: English ⇨ Hebrew

I wash the car

אני רוחץ את
(אני רוחץ את the car)

Google translate

I wash the floor

אני שוטפת את
(אני שוטפת את the floor)

Google translate
Example: English 🇬🇧 Hebrew

I wash the car

אני רוחץ את

(I wash the car)

I wash the floor

אני שוטפת את

(I wash the floor)
Example: English 🆕 Hebrew

I wash the car
אני רוחץ את
(אֶנְי רוּחַץ אַת
dike  the car)

I wash the floor
אני שוטפת את
(אֶנְי שוּפוּפַת אַת
dike the floor)

Masculine
Feminine
Hebrew Verbs are morphologically marked for Gender
(and Number, and Person, and Tense...)
Example: English ☚ Hebrew

I wash the car

אני רוחץ את
(1 וושן the car)

Masculine

I wash the floor

אני שוטפת את
(1 פשטי the floor)

Feminine

Are these bad translations?
Example: English ➔ Hebrew

I wash the floor
אני שוטפת את הרצפה

I wash the car
אני רוחץ את מכונית

No. These are actually quite good.

No gender information in source. Target must indicate gender. ♂ translator uses world knowledge.

Are these bad translations?
Let’s have some fun
Language Models as Social Indicators

• I love her
• I love him
Language Models as Social Indicators

- I love her
- I love him

אני אוהב אתיה
אני אוהבת אתיה
Language Models as Social Indicators

- I love her
- I love him
- I love meat
- I love vegetables

• אני אוהב אתה
• אני אוהבת אתו
Language Models as Social Indicators

- I love her
- I love him
- I love meat
- I love vegetables

אני אוהב אתה
אני אוהבת אתו
אני אוהב בשר
אני אוהבת ירקות
Language Models as Social Indicators

- I love her
- I love him
- I love meat
- I love vegetables
- I love to eat
- I love to cook

אני אוהב אווה
אני אוהב אווה
אני אוהב בשר
אני אוהב ירקות
אני אוהב אוכל
אני אוהב לעקוב
Language Models as Social Indicators

- I love her
- I love him
- I love meat
- I love vegetables
- I love to eat
- I love to cook

אני אוהב אתה
אני אוהבת אוטו
אני אוהב ובשר
אני אוהבת ירקות
אני אוהב לאכול
אני אוהבת לבשל
Language Models as Social Indicators

- I love her
- I love him
- I love meat
- I love vegetables
- I love to eat
- I love to cook
- I love hash
- I love marijuana

אני אוהב אתנה
אני אוהב אתו
אני אוהב בשר
אני אוהבת ירקות
אני אוהב לאכול
אני אוהבת לבשל
אני אוהב קנאביס
Language Models as Social Indicators

- I love her
- I love him
- I love meat
- I love vegetables
- I love to eat
- I love to cook
- I love hash
- I love marijuana
- אני אוחב אתיה
- אני אוחבת אותו
- אני אוחב בשר
- אני אוחבת ירקות
- אני אוחב לאכל
- אני אוחבת לבלש
- אני אוחבخشיש
- אני אוחבת מריחואנה
Language Models as Social Indicators

- I hate him.
- אני שונאת אותו.
Language Models as Social Indicators

- I hate him.
- I hate her.
- אני שונאת אתו.
Language Models as Social Indicators

- I hate him.
- I hate her.

אני שונאת אותו.
אני שונאת אשה.
Language Models as Social Indicators

- I hate him.
- I hate her.
- I hate him
Language Models as Social Indicators

• I hate him.
• I hate her.
• I hate him

Really? A dot?! Not very stable...

אני שונאת אתו.
אני שונאת אתה.
אני שונא אתו.
Language Models as Social Indicators

- I hate him.
- I hate her.
- I hate him
- I hate her

Really? A dot?!
Not very stable...
Language Models as Social Indicators

- I hate him.
- I hate her.
- I hate him.
- I hate her.

Really? A dot?!
Not very stable...

Hmm... is there a message here after all?

אני שונאת אתו.
אני שונאת אתנה.
אני שונא אתו.
אני שונאת אתנה.

•
•
•
•
Language Models as Social Indicators

- I love
- I hate
Language Models as Social Indicators

- I love
- I hate
Back to Machine Translation
One English ⇄ Many Hebrew

love

Let's assume this is solved
One English ↔ Many Hebrew

love

Aḥavat Aḥobim Aḥob

Need to acquire more knowledge
One English ❭ Many Hebrew

love

أشבל אוחבים

Need to acquire more knowledge

• ... use larger parallel corpora
One English ⇔ Many Hebrew

Need to acquire more knowledge
- ... use larger parallel corpora
- ... use dictionaries

Let's assume this is solved.
One English ??? Many Hebrew

Need to acquire more knowledge
• ... use larger parallel corpora
• ... use dictionaries
• ... use FSAs to model inflections

LET'S ASSUME THIS IS SOLVED
One English ??? Many Hebrew

Need to acquire more knowledge

• ... use larger parallel corpora
• ... use dictionaries
• ... use FSAs to model inflections
• Let’s assume this is solved
One English ⇔ Many Hebrew

Hebrew ⇔ English:
easy
אני אוהבת אנני
I love
I love
אנחנו אוהבים
We love
We love
Don't worry about it.
Just say "love".
The reader will decide.
Which form to choose?
Translator must decide.
HOW??
One English ⇔ Many Hebrew

Hebrew ⇔ English:
easy

I love

love

Don't worry about it. Just say "love". The reader will decide. Which form to choose? Translator must decide. How??
One English ⇔ Many Hebrew

Hebrew ⇔ English:

- easy
  - אהבת אני
  - I love
- love
  - אוהבים אנו
  - We love
- loves
  - אוהבים אנו
  - We love

Don’t worry about it. Just say “love”. The reader will decide.

Which form to choose? Translator must decide.
One English ↔ Many Hebrew

Hebrew ↔ English: easy

אני אוהב אני
I love

We love

Don’t worry about it. Just say “love”. The reader will decide.

Hebrew ↔ English: hard

אני אוהב אני
I love

We

אני אוהב אני
I love

We

Don’t worry about it. Just say “love”. The reader will decide.
One English ⇔ Many Hebrew

Hebrew ⇔ English: easy

I love

Hebrew ⇔ English: hard

Don’t worry about it. Just say “love”. The reader will decide.

Which form to choose? Translator must decide. HOW??
When translating into an MRL:

- Many possible word forms
  - Hard to acquire [but assume its solved]
- Need to choose correct inflection
One English ?? Many Hebrew

Hebrew ☐ English: easy

אני אוהב אני ☐ I love

We love ☐

Don’t worry about it.
Just say “love”.
The reader will decide.

Which form to choose?
Translator must decide.

HOW??

Hebrew ☐ English: hard

אני אוהבת אני ☐ I love

We love ☐

אני אוהבות אני ☐ We love

Which form to choose?
Translator must decide.

HOW??

Don’t worry about it.
Just say “love”.
The reader will decide.
Don’t worry about it. Just say “love”. The reader will decide.

Which form to choose? Translator must decide.

Just choose one at random? In the worst case we’ll insult someone..
Hebrew Verbs agree with Subject on gender and number
Agreement dictates form

אני אוהב
אני אוהבת
אני אוהב
אני אוהבות
אני אוהב
אני אוהבות

Agreement dictates form

אני אוהב

אני אוהבת

אני אוהבتم

אני אוהבות
Agreement dictates form

אני אוהב

אני אוהבת

אני אוהבים

אני אוהבות

I love
Agreement dictates form

I love

אני אוהב

singular

אני אוהבת

singular singular

אני אוהבות

singular singular

אני אוהבים

plural singular

אני אוהבות

plural singular
The girls love

בוחרות

The girls love plural fem

حبות אורות

The girls love plural masc

حبות אורות

The girls love plural fem

حبות אורות
<table>
<thead>
<tr>
<th>English</th>
<th>Hebrew</th>
</tr>
</thead>
<tbody>
<tr>
<td>The girls love</td>
<td>הבנות אחזות</td>
</tr>
<tr>
<td></td>
<td>הבנות אוהבות</td>
</tr>
<tr>
<td></td>
<td>הבנות אוהבים</td>
</tr>
</tbody>
</table>

no missing information
The girls love...
When translating into an MRL:

· Many possible word forms
  - Hard to acquire [but assume its solved]
· Need to choose correct inflection
· Inflection is determined based on information which is external to the word
The boy washes the car

The girl washes the car
The boy washes the car

The girl washes the car
The boy washes the car
The girl washes the car
The boy washes the car

The girl washes the car

Good job Franz!
The boy washes the car

The girl washes the car

نتער רוחץ את המכונית

הילדה רוחצת את המכונית

Good job Franz?
The boy washes the car

The girl washes the car

Good job Franz?
The boy with the sunglasses washes the floor.

The girl with the sunglasses washes the car.

Good job Franz?
The boy with the sunglasses washes the floor.

The girl with the sunglasses washes the car.

Good job Franz?
The boy with the sunglasses washes the floor.

The girl with the sunglasses washes the car.

Good job Franz?
The boy with the sunglasses washes the floor.

The girl with the sunglasses washes the car.

Good job Franz?
What happened?

- Long distance agreement
- Can’t be represented in phrase-table
- Can’t be represented in n-gram LM
  - Local “semantic” information from LM/Phrase
  - Bad translation (ungrammatical)
What happened?

- Long distance agreement
- Can’t be represented in phrase-table
- Can’t be represented in n-gram LM
  - Local “semantic” information from LM/Phrase
  - Bad translation (ungrammatical)

It’s not Franz’s fault, but the system’s
When translating into an MRL:

- Many possible word forms
  - Hard to acquire [but I assume its solved]
- Need to choose correct inflection
- Inflection is determined based on information which is external to the word and frequently far away from it
When translating into an MRL:

- Many possible word forms
  - Hard to acquire [but I assume its solved]
- Need to choose correct inflection
  - Inflection is determined based on information which is external to the word and frequently far away from it

### Distance from Verb to Subject in the Hebrew Dependency Treebank (news domain)

<table>
<thead>
<tr>
<th>S-V Dep-Length</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3218</td>
<td>42%</td>
</tr>
<tr>
<td>2</td>
<td>1504</td>
<td>19%</td>
</tr>
<tr>
<td>3</td>
<td>914</td>
<td>12%</td>
</tr>
<tr>
<td>4</td>
<td>405</td>
<td>5%</td>
</tr>
<tr>
<td>5</td>
<td>297</td>
<td>4%</td>
</tr>
<tr>
<td>&gt;5</td>
<td>1322</td>
<td>17%</td>
</tr>
</tbody>
</table>

Distance from Verb to Subject in the Hebrew Dependency Treebank (news domain)
When translating into an MRL:

- Many possible word forms
  - Hard to acquire [but I assume it's solved]
- Need to choose correct inflection
  - Inflection is determined based on information which is external to the word and frequently far away from it

2 words apart is already though for reliably estimating in an n-gram based system!

<table>
<thead>
<tr>
<th>S-V Dep-Length</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3218</td>
<td>42%</td>
</tr>
<tr>
<td>2</td>
<td>1504</td>
<td>19%</td>
</tr>
<tr>
<td>3</td>
<td>914</td>
<td>12%</td>
</tr>
<tr>
<td>4</td>
<td>405</td>
<td>5%</td>
</tr>
<tr>
<td>5</td>
<td>297</td>
<td>4%</td>
</tr>
<tr>
<td>&gt;5</td>
<td>1322</td>
<td>17%</td>
</tr>
</tbody>
</table>

Distance from Verb to Subject in the Hebrew Dependency Treebank (news domain)
When translating into an MRL:

- Many possible word forms
  - Hard to acquire [but I assume its solved]
- Need to choose correct inflection
- Inflection is determined based on information which is external to the word and frequently far away from it
When translating into an MRL:

- Many possible word forms
  - Hard to acquire [but I assume it's solved]
- Need to choose correct inflection
- Inflection is determined based on information which is *external* to the word and frequently far away from it

❓ Phrase based + N-gram LM can't do it
What if both languages are MRLs?

- Gender/number marked on both sides
- No need for word-external information
- We can translate word→word again
- MRL → MRL is easy!
What if both languages are MRLs?

- Gender/number marked on both sides
- No need for word-external information
- We can translate word→word again
- MRL → MRL is easy!

Wrong!
What if both languages are MRLs?

- Gender/number marked on both sides
- **But:**
  - agreement patterns differ between languages
  - gender information differs between languages
What if both languages are MRLs?

Example:
- Spanish and Hebrew have adjective-noun agreement
What if both languages are MRLs?

Example:

- Spanish and Hebrew have **adjective-noun** agreement

  - new shirt
    - חולצה חדש
    - nueva camisa
What if both languages are MRLs?

Example:

- Spanish and Hebrew have **adjective-noun** agreement

  - new shirt
    - חולצת חדש
    - nueva camisa

  - new car
    - מכונית חדשה
    - nuevo automovil
What if both languages are MRLs?

Example:
- Spanish and Hebrew have **adjective-noun** agreement
  - new shirt
    - חולצת חדשה
    - nueva camisa
  - new car
    - מכונית חדשה
    - nuevo automóvil
  - new computer
    - מחשב חדש
    - nueva computadora
What if both languages are MRLs?

Example:
- Spanish and Hebrew have **adjective-noun** agreement

  - new shirt
    - חולצה חדש
    - nueva camisa

  - new car
    - מכונית חדש
    - nuevo automovil

  - new computer
    - מחשב חדש
    - nueva computadora
What if both languages are MRLs?

Example:
- Many-to-many mapping
- Correct form still depends on external information
- More chances for error
- Acquiring all the pairs from parallel corpora is harder

- new shirt
  • חולצה חדשה
  • nueva camisa

- new car
  • מכונית חדשה
  • nuevo automovil

- new computer
  • מחשב חדש
  • nueva computadora

- novo automovil
  • מחשב חדש
  • nueva
Phrase-tables and n-grams still can’t do it

Must consider (at least) syntax
When Translating into an MRL:

• MT systems must be aware of gender/number
• Should have a notion of agreement
• Use syntax to enforce agreement
The boy with the sunglasses washes the floor.
The boy with the sunglasses washes the floor.
The boy with the sunglasses washes the floor.
The boy with the sunglasses washes the floor.

Source-side Syntax:

- **subj**: The boy
- **with**: the sunglasses
- **obj**: the floor

Agreement:

- **masc/sing**: נער
- **fem /sing**: ילד
- **masc/plural**: בני
- **fem /plural**: אחיות

The pronoun agreement is marked on the diagram.
The boy with the sunglasses washes the floor.

The boy

The

with

sunglasses

the

floor

masc/sing

masc/sing

masc/sing

masc/plural

fem/sing

fem/plural

Agree!
Source-side Syntax

Problems:
- How to obtain gender/number information?
- How to decode efficiently?
- Agreement behavior is not always that simple
Target-side Syntax

xLNT transducers can model agreement
Target-side Syntax

**xLNT transducers can model agreement**

NNmasc/sing (ילד) \(\rightarrow\) boy
VBmasc/sing (שותף) \(\rightarrow\) washes
VBfem/sing (שותפת) \(\rightarrow\) washes

NPmasc/sing([x0:DT x1:NNmasc/sing x2:PP] \(\rightarrow\) x0 x1 x2)
VPmasc/sing (x0:VB masc/sing x1:NP) \(\rightarrow\) x0 x1
S masc/sing(x0:NPmasc/sing x1:VPmasc/sing) \(\rightarrow\) x0 x1
Target-side Syntax

xLNT transducers can model agreement

NNmasc/sing (ילד)  boy
VBmasc/sing (שוטף)  washes
VBfem/sing (שוטפת)  washes

NPmasc/sing(x0:DT x1:NNmasc/sing x2:PP)
□ x0 x1 x2
VPmasc/sing (x0:VB masc/sing x1:NP)
□ x0 x1
S masc/sing(x0:NPmasc/sing x1:VPmasc/sing)
□ x0 x1

Gender and number information encoded in the lexical rules
Target-side Syntax

xLNT transducers can model agreement

Gender and number information encoded in the lexical rules

Agreement information encoded in the grammar

NNmasc/sing (ילד) boy
VBmasc/sing (שוטף) washes
VBfem/sing (שוטפת) washes

NPmasc/sing(x0:DT x1:NNmasc/sing x2:PP)  □ x0 x1 x2
VPmasc/sing (x0:VB masc/sing x1:NP)  □ x0 x1
Smasc/sing(x0:NPmasc/sing x1:VPmasc/sing)  □ x0 x1
Target-side Syntax

xLNT transducers can model agreement

Problems:
- How to obtain gender/number information?
- Grammar is going to be huge (can we make it smaller?)
- How are we going to obtain the grammar?

efficiently encoding morphological processes in a treebank grammar ? an open research question
On The Parsing Side of Things

- Most work on parsing MRLs:
  - consider morphology to be a lexicon-level issue
    - Many inflections → high OOV rate
  - Ignoring morphology at syntax-level
  - PCFGLA works frustratingly well
    - Modest benefits to parsing accuracy
      - PCFGLA still better 😁
On The Parsing Side of Things

- Most work on parsing MRLs:
  - consider morphology to be a lexicon-level issue
    - Many inflections → high OOV rate
  - Ignoring morphology at syntax-level
    - PCFGLA works frustratingly well
- Modest benefits to parsing accuracy
- PCFGLA still better
On The Parsing Side of Things

• Most work on parsing MRLs:
  - consider morphology to be a lexicon-level issue
    • Many inflections → high OOV rate
  - Ignoring morphology at syntax-level
  - PCFGLA works frustratingly well

• Recently:
  → Modest benefits to parsing accuracy
  → PCFGLA still better
On The Parsing Side of Things

- Most work on parsing MRLs:
  - consider morphology to be a lexicon-level issue
    - Many inflections → high OOV rate
  - Ignoring morphology at syntax-level
  - **PCFGLA works frustratingly well**

- Recently:
  - Smarter modeling of morphology at syntax level

→ **Modest benefits to parsing accuracy**
→ **PCFGLA still better**
On The Parsing Side of Things

• Most work on parsing MRLs:
  - consider morphology to be a lexicon-level issue
    • Many inflections → high OOV rate
  - Ignoring morphology at syntax-level
  - **PCFGLA works frustratingly well**

• Recently:
  - Smarter modeling of morphology at syntax level
  - Using morphological agreement to improve

→ **Modest benefits to parsing accuracy**
→ **PCFGLA still better**
On The Parsing Side of Things

• Most work on parsing MRLs:
  - consider morphology to be a lexicon-level issue
    • Many inflections → high OOV rate
  - Ignoring morphology at syntax-level
  - **PCFGLA works frustratingly well**

• Recently:
  - Smarter modeling of morphology at syntax level
  - Using morphological agreement to improve parsing
    → **Modest benefits to parsing accuracy**
    → **PCFGLA still better**
On The Parsing Side of Things

- Most work on parsing MRLs:
  - consider morphology to be a lexicon-level issue
  - Many inflections → high OOV rates
  - Ignoring morphology at syntax-level
  - **PCFGLA works frustratingly well**

- Recently:
  - Smarter modeling of morphology at syntax level
  - Using morphological agreement to improve parsing
  → **Modest benefits to parsing accuracy**
  → **PCFGLA still better**
Rebels without a cause?

- Syntax-based MT:
  - Neat!
  - Only marginally better than phrase-based
Rebels without a cause?

- Syntax-based MT: Neat!
- Only marginally better than phrase-based

English grammaticality is relatively easy to capture using local information
Rebels without a cause?

- Syntax-based MT:
  - Neat!
  - Only marginally better than phrase-based

- Syntax-level morphology in parsing:
  - Neat!
  - Only marginally better than ignoring it

English grammaticality is relatively easy to capture using local information.
Rebels without a cause?

• Syntax-based MT:
  - Neat!
  - Only marginally better than phrase-based

• Syntax-level morphology in parsing:
  - Neat!
  - Only marginally better than ignoring it

English grammaticality is relatively easy to capture using local information

• I should work harder.
• Not many agreement mistakes to begin with.
• Agreement is a generation issue more than a parsing one.

• I should work harder.
• Not many agreement mistakes to begin with.
• Agreement is a generation issue more than a parsing one.
Rebels without a cause?

• Syntax-based MT:
  - Neat!
  - Only marginally better than phrase-based

• Syntax-level morphology in parsing:
  - Neat!
  - Only marginally better than ignoring it

Both are crucial for translating into MRLs!
To Conclude

- Translating into MRLs brings new challenges

- Syntax is crucial
  - If you are not looking into syntax, you should!
  - If you are looking into syntax – look deeper!

- Plenty of interesting work to be done
To Conclude

• Translating into MRLs brings new challenges

• Syntax is crucial
  - If you are not looking into syntax, you should!
  - If you are looking into syntax – look deeper!

• Plenty of interesting work to be done
  - Finishing up my phd on parsing
  Looking for a postdoc position for next year