integrated morphology for translation

kevin knight
university of southern california

MTMRL / Haifa / January 2011
Automatic Language Translation
Automatic Language Translation

is darn hard
Even for simple sentences...

Input:
贝尔当场死亡，他的两个朋友受伤。

Correct:
Bell died on the spot and his two friends were injured.

MT output:
Bell died on the spot and injured his two friends.
The government of the United States today listed polar bears as a species on the verge of danger.
Need Lots of Knowledge

• What do words mean?
• What is the relation between meanings of words and the meaning of a sentence?
• What makes a sentence grammatical?
• Why does the same word have so many forms?
• How do those forms play a role in translation?
Brown et al., 1988

Translation Model Tables

Align

Target Language Model Tables

French

English
Brown et al, 1988

words fly around

Decode

Translation Model Tables

Target Language Model Tables

French

English
Brown et al, 1988

words fly around

Decode

Translation Model Tables
Target Language Model Tables

Align

French English
Brown et al, 1988

eats       mange
ate        mangé
eating     manger
eat         manger
...
Brown et al, TMI 1992
Brown et al, TMI 1992

The diagram illustrates the process of translation and aligning language models. Words fly around the system, going through the processes of decoding, translation, and modeling. The intermediate languages of French and English are involved, with analysis and synthesis steps. The diagram highlights the flow of information and the use of model tables in the translation process.
Brown et al, TMI 1992

Translation Model Tables

Target Language Model Tables

words fly around

linguistics free!

Intermediate French

Intermediate English

 Align

Decode

analyze

analyze

synthesize

French

English
Brown et al, TMI 1992

Peter Brown: “We gotta take the languages and line ’em up.”

Je ne sais pas. | Je vous le donnerai.
⇒ Je sais ne pas. ⇒ Je donnerai le_DPRO vous_IPRO.

Why should farmers be growing less wheat?
⇒ Why farmers should be growing less wheat QINV

Because of errors in grammatical tagging, compounded with the primitive nature of the rules that we employ to achieve this goal, we succeed only about 40% of the time.

Mangez-vous des légumes?
⇒ Vous mangez des légumes QINV1
He was eating the peas more quickly than I.

⇒ He PAST_PROGRESSIVE to_eat the pea N_PLURAL quick er ADV than I.

Ils se sont lavés les mains sales.

⇒ Ils 3RD_PERSON_PLURAL_PAST laver se_RPRO les sale main N_PLURAL.

Notice in the last example that we retain no indication of the original number on French adjectives. We also discard any distinction in gender. Thus, in the intermediate French, adjectives always appear in their masculine singular form.
We assign senses to 1000 of the most frequent French words. For example, we map *prendre* to *prendre_1* in the sentence

*Je vais prendre ma propre voiture,*

but to *prendre_2* in the sentence

*Je vais prendre ma propre décision.*
Brown et al, TMI 1992

We restrict our attention to vocabularies of 40,809 English words and 57,802 French words. In the enhanced system, morphological analysis reduces these to 33,041 English morphemes and 31,115 French morphemes.

We estimated the parameters of the translation model for each system from a set of 1,778,620 pairs of French and English sentences from the Canadian Hansard data [1, 2]. Each of these sentences is 30 words or less in length. We tested both systems on the same set of 100 randomly selected Hansard sentences each containing at most 10 words. We judged as acceptable 39 of the translations produced by the simpler system as compared with 60 of those produced by the enhanced system.
Brown et al, TMI 1992

In work of this type, it is desirable to be able ascribe certain increments of performance to certain of the steps in the analysis or synthesis component, and thus to assess the value of the various transformations. Making such an assessment would require of us that we construct a series of analysis and synthesis components with different members of the series including different ones of the steps that make up the complete system. Unfortunately, each such construction must have a differently trained statistical transfer component. Because training is a costly undertaking, we have not made any of these collateral investigations and are, therefore, unable to say which of the new analysis and synthesis steps is the most valuable.
Brown et al, TMI 1992

In work of this type, it is desirable to be able ascribe certain increments of performance to certain of the steps in the analysis or synthesis component, and thus to assess the value of the various transformations. Making such an assessment would require of us that we construct a series of analysis and synthesis components with different members of the series including different ones of the steps that make up the complete system. Unfortunately, each such construction must have a differently trained statistical analyst, and the construction is a costly undertaking, we have not made any of these comparisons, and so unable to say which of the new analysis and synthesis steps is of more value.
Brown et al, TMI 1992

In work of this type, it is desirable to be able as a first approximation to choose the components of the analysis or synthesis components with different members of the various transformations. Making such an assessment is a costly undertaking, we have not made any of these combinations of analysis and synthesis steps. Unfortunately, each such construction must have a differently trained statistical construction. It is therefore impossible to say which of the new analysis and synthesis steps that make up the complete system. Unfortunately, each such construction must have a differently trained statistical construction. It is therefore impossible to say which of the new analysis and synthesis steps that make up the complete system.
Berger et al, 1996

Baseball managers (who rank among the better paid statistical modelers) employ batting averages, compiled from a history of at-bats, to gauge the likelihood that a player will succeed in his next appearance at the plate. Thus informed, they manipulate their lineups accordingly. Wall Street speculators (who rank among the best paid statistical modelers) build models based on past stock price movements to predict tomorrow’s fluctuations and alter their portfolios to capitalize on the predicted future. At the other end of the pay scale reside natural language researchers, who design language and acoustic models for use in speech recognition systems and related applications.

The natural language processing community has made significant progress toward increasing the
Berger et al, 1996

...to make predictions about the future behavior of the process.

Baseball managers (who rank among the best) employ batting averages, compiled from a history of all past at-bats, to predict that a player will succeed in his next appearance at the plate and manipulate their lineups accordingly. Wall Street speculators (who are greatly paid statistical modelers) build models based on past price movements to forecast tomorrow’s fluctuations and alter their portfolios to capitalize on the predicted future. At the other end of the pay scale reside natural language researchers, who design language and acoustic models for use in speech recognition systems and related applications.

The nature of language is an area of persistent progress toward increasing the
Renaissance Technologies LLC, one of the most successful hedge-fund companies ever... Medallion fund averaged returns of about 45% a year, after fees, since its inception in 1988.

Now [new co-CEOs] Peter Brown and Bob Mercer ... must steer the firm through challenging waters.
Brown et al, 2010

Wall Street Journal, March 16, 2010

Renaissance Technologies LLC, one of the most successful hedge-fund companies ever... Medallion fund averaged returns of about 45% a year, after fees, since its inception in 1988.

Now [new co-CEOs] Peter Brown and Bob Mercer ... must steer the firm through challenging waters.

<table>
<thead>
<tr>
<th>Year</th>
<th>Medallion Fund Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>$100</td>
</tr>
<tr>
<td>1989</td>
<td>145</td>
</tr>
<tr>
<td>1990</td>
<td>210</td>
</tr>
<tr>
<td>1991</td>
<td>305</td>
</tr>
<tr>
<td>1992</td>
<td>442</td>
</tr>
<tr>
<td>1993</td>
<td>641</td>
</tr>
<tr>
<td>1994</td>
<td>929</td>
</tr>
<tr>
<td>1995</td>
<td>1,348</td>
</tr>
<tr>
<td>1996</td>
<td>1,954</td>
</tr>
<tr>
<td>1997</td>
<td>2,833</td>
</tr>
<tr>
<td>1998</td>
<td>4,108</td>
</tr>
<tr>
<td>1999</td>
<td>5,957</td>
</tr>
<tr>
<td>2000</td>
<td>8,638</td>
</tr>
<tr>
<td>2001</td>
<td>12,525</td>
</tr>
<tr>
<td>2002</td>
<td>18,162</td>
</tr>
<tr>
<td>2003</td>
<td>26,334</td>
</tr>
<tr>
<td>2004</td>
<td>38,185</td>
</tr>
<tr>
<td>2005</td>
<td>55,368</td>
</tr>
<tr>
<td>2006</td>
<td>80,283</td>
</tr>
<tr>
<td>2007</td>
<td>116,410</td>
</tr>
<tr>
<td>2008</td>
<td>168,795</td>
</tr>
<tr>
<td>2009</td>
<td>244,753</td>
</tr>
</tbody>
</table>
1992-2010: Rising Ambitions

• **Large parallel data!**
  – Eventually observe 99.9% of all word forms that will ever actually occur (& their translations)?

• **Large target language models!**
  – Output “casa roja” or “casa rojo”? 

• **Lots of language pairs!**
  – Too much work to “line up” every pair?

• **Avoid medicine with bad side-effects!**
  – What if analyzer/synthesizer makes mistakes, and what if they mess up what was already working fine?
Large Language Models

The weather is cool.

→ El clima es fresco.

(el clima = 6m web hits, la clima = 228k)

That's cool.

→ Eso es genial.

Can you dig it?

→ ¿Puedes creerlo? (i.e., can you believe it?)

I dig her.

→ Yo la excavación. (i.e., I-her-excavation)
Live News Translation

English translation

which touched several schools and villages, a number of victims of the first ever major [...]

Foreign language speech recognition

Foreign language speech recognition

searchable archive

searchable archive

news broadcast

news broadcast

news broadcast
Online Help

Procesadores

Aplicación de material de interfaz térmica para procesadores de desktop

La corriente soluciones térmicas que acompañan a los procesadores Intel® Desktop y servidores incluyen un nuevo material de interfaz térmica aplicado en la parte inferior del disipador térmico mediante una aplicación de 3 barras que se realiza en la fábrica (figura 1). Este material de interfaz térmica en el disipador térmico del ventilador asegura que una transferencia de calor adecuada tiene lugar entre el difusor térmico integrado del procesador y el disipador térmico del ventilador.
User-Generated Content

TripAdvisor.com: Millions of English user reviews translated into French, Spanish, German, etc.

"Uno de los mejores sitios de sushi en el área de los Ángeles"
Sushi Katsu-ya

1/1 considera esta crítica muy útil

Este es un lugar favorito de almuerzo de la mía. La ubicación es estupenda Sherman Oaks también, pero yo preferiría un poco este, quizás porque es tu clásico en el un-strip-sushi-gourmet de centro comercial. Como tal, presume la decoración mismo no de la decoración, pero ¿qué importa? Lujoso lugares de sushi son normalmente no muy bueno (algunas excepciones, pero en la verdad demasiado a menudo). No te pierdas los bollos de cangrejo o el arroz crujiente, pero es todo bien. El servicio es amable, pero está siempre lleno, reservarlas, reservarlas, reservarlas!

Esta crítica es la opinión subjetiva de un miembro de TripAdvisor,
# Deployed Statistical MT Systems

<table>
<thead>
<tr>
<th>Region</th>
<th>Company X</th>
<th>Company Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western Europe</td>
<td>Danish, Dutch, Finnish, French, German, Greek, Italian, Norwegian, Portuguese, Spanish, Swedish</td>
<td>Catalan, Danish, Dutch, French, Galician, German, Greek, Icelandic, Irish, Italian, Maltese, Norwegian, Polish, Portuguese, Spanish, Swedish, Welsh</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>Bulgarian, Czech, Hungarian, Polish, Romanian, Russian, Serbian, Turkish</td>
<td>Albanian, Belarusian, Bulgarian, Croatian, Czech, Estonian, Finnish, Hungarian, Latvian, Lithuanian, Macedonian, Romanian, Russian, Serbian, Slovak, Slovenian, Turkish, Ukrainian, Yiddish</td>
</tr>
<tr>
<td>Middle East &amp; Africa</td>
<td>Arabic, Hausa, Hebrew, Pashto, Persian, Somali, Urdu</td>
<td>Afrikaans, Arabic, Hebrew, Persian, Swahili</td>
</tr>
<tr>
<td>Asia</td>
<td>Chinese, Hindi, Indonesian, Japanese, Korean, Thai</td>
<td>Chinese, Filipino, Hindi, Indonesian, Japanese, Korean, Malay, Thai</td>
</tr>
<tr>
<td>English source</td>
<td>Correct Swahili</td>
<td>MT Swahili</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>I am reading a book.</td>
<td>Ninasoma kitabu.</td>
<td>I am kusoma kitabu.</td>
</tr>
<tr>
<td></td>
<td>I-PRESENT-read book</td>
<td></td>
</tr>
<tr>
<td>You are reading a book.</td>
<td>Unasoma kitabu.</td>
<td>Wewe ni kusoma kitabu.</td>
</tr>
<tr>
<td></td>
<td>you-PRESENT-read book</td>
<td></td>
</tr>
<tr>
<td>He is reading a book.</td>
<td>Anasoma kitabu.</td>
<td>Yeye ni kusoma kitabu.</td>
</tr>
<tr>
<td></td>
<td>he-PRESENT-read book</td>
<td></td>
</tr>
</tbody>
</table>
African Languages
African Languages
African Languages
African Languages
African Languages
African Languages
African Languages
African Languages
Almost as big as the Moon and even more languages spoken!
Even if you don’t care about that

• Translation into
  • Czech
  • Russian
  • Japanese
  • German
  • Korean
  • Hebrew
  • Arabic

is problematic!
Even into Spanish

Gender agreement

<table>
<thead>
<tr>
<th>English input</th>
<th>Spanish MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red house</td>
<td>Casa roja</td>
</tr>
<tr>
<td>Yellow house</td>
<td>Casa amarilla</td>
</tr>
<tr>
<td>Black house</td>
<td>Casa negro</td>
</tr>
<tr>
<td>Orange house</td>
<td>Naranja casa</td>
</tr>
<tr>
<td>An orange house</td>
<td>Una casa de color naranja</td>
</tr>
</tbody>
</table>

Verb inflection

<table>
<thead>
<tr>
<th>English input</th>
<th>Spanish MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>I saluted the flag.</td>
<td>Saludé a la bandera.</td>
</tr>
<tr>
<td>You saluted the flag.</td>
<td>Usted saludó a la bandera.</td>
</tr>
<tr>
<td>He saluted the flag.</td>
<td>Saludó a la bandera.</td>
</tr>
<tr>
<td>She saluted the flag.</td>
<td>Rindió homenaje a la bandera.</td>
</tr>
<tr>
<td>We saluted the flag.</td>
<td><strong>Saludamos</strong> la bandera.</td>
</tr>
<tr>
<td>They saluted the flag.</td>
<td>Se saludaron a la bandera.</td>
</tr>
</tbody>
</table>
Recent Research

- **Source ➔ Source-prime**
  - Analyzer re-orders source language
  - Makes use of source parsing & morphology
    - Not available in 1992!
  - Can encode uncertainties in source lattice

- **Target ➔ Target-prime**
  - Translate into target lemma sequences
  - Synthesizer guesses inflections
    - Which may have no correlate in English anyway

- Factored Models
Recent Research

• Underlying engine is often still linguistics-free
  – Words fly around at decoding time
  – “Distortion” cost
  – Target language model supposed to sort it out

• In other words:
  – We already have a linguistics-free, distortion-based MT system
  – Let’s bolt on some syntax/morph without disturbing the engine
Still, Lots of Details ...

北极熊 ➔ polar bear
北极熊 ➔ polar bears

<table>
<thead>
<tr>
<th>Chinese-to-English, 200m words training</th>
<th>Test Bleu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline state-of-the-art MT</td>
<td>32.0</td>
</tr>
</tbody>
</table>
| **1) Separate** English affixes in training data  
2) Train, decode  
3) Re-join affixes in decoder output | 30.6      |
| **1) Remove** English affixes in training data  
2) Train, decode  
3) Guess inflections for decoder output | 31.6      |
| 1) Decode  
2) Remove and re-generate affixes in decoder output | 31.4      |
Not Super-Satisfying?

• Want rules like “Translate direct object by moving it before the verb and marking it with an accusative affix”
• We don’t have such morphological processes integrated in
• Syntactic movement has been getting attention, though
  – Next: description of syntax-based MT
  – Then: possible directions for morphology
Syntax-Based
SMT

这种现象在寒冷的冬季尤其明显。
这种现象在寒冷的冬季尤其明显。
这种现象在寒冷的冬季尤其明显。
Tree-Based Output

枪手 被 警方 击毙．
The gunman killed by police.

Decoder
Hypothesis #1

枪手 被 警方 击毙．
Gunman by police shot.
The gunman was killed by police.

How does a Chinese string become an English tree, or vice-versa?
Top-Down Tree Transducer

Original input:

Transformation:
Top-Down Tree Transducer

Original input:

Transformation:
Top-Down Tree Transducer

Original input:

Transformation:
Top-Down Tree Transducer

Original input:
he enjoys listening to music

Transformation:
s NP
  | PRO
  | he
  |
  VP
    |
    listening
to music
Top-Down Tree Transducer

Original input:

Transformation:
Top-Down Tree Transducer

Original input: he enjoys listening to music

Final output: kare wa, ongaku o kiku, no, ga, daisuki, desu

To get total probability, multiply probabilities of the individual steps.
Top-Down Tree Transducer

• Introduced by Rounds (1970) & Thatcher (1970)
  “... parts of mathematical linguistics can be formalized easily in a tree-automaton setting ...” (Rounds 1970, “Mappings on Grammars and Trees”, Math. Systems Theory 4(3))

• Large theory literature
  – e.g., Gécseg & Steinby (1984), Comon et al (1997)

• Once again re-connecting with NLP practice
  – e.g., Knight & Graehl (2005), Knight (2007), May & Knight (2006), Maletti (2010), ATANLP workshop at ACL 2010, etc.
Tree Transducers Can be Extracted from Bilingual Data (Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

VBD(felt) → 有
VBN(obliged) → 责任
VB(do) → 尽
NN(part) → 一份
NN(part) → 一份力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C) → x0
S(x0:NP-C x1:VP) → x0 x1
Sample Subject-Verb-Object Rules

CHINESE / ENGLISH

0.82 \( S(x_0:NP-C\ VP(x_1:VBD\ x_2:NP-C)\ x_3:.) \rightarrow x_0\ x_1\ x_2\ x_3 \)
0.02 \( S(x_0:NP-C\ VP(x_1:VBD\ x_2:NP-C)\ x_3:.) \rightarrow x_0\ x_1\ ,\ x_2\ x_3 \)
0.01 \( S(x_0:NP-C\ VP(x_1:VBD\ x_2:NP-C)\ x_3:.) \rightarrow x_0\ ,\ x_1\ x_2\ x_3 \)

ARABIC / ENGLISH

0.54 \( S(x_0:NP-C\ VP(x_1:VBD\ x_2:NP-C)\ x_3:.) \rightarrow x_0\ x_1\ x_2\ x_3 \)
0.44 \( S(x_0:NP-C\ VP(x_1:VBD\ x_2:NP-C)\ x_3:.) \rightarrow x_1\ x_0\ x_2\ x_3 \)
Decoding

- argmax $P(\text{etree} \mid \text{cstring})$
  etree

- Difficult search problem
  - Bottom-up CKY parser
  - Builds English constituents on top of Chinese spans
  - Record of rule applications (the derivation) provides information to construct English tree
  - Returns k-best trees
Syntax-Based Decoding

Rules apply when their right-hand sides (RHS) match some portion of the input.
This 7 people include astronauts from France and Russia.
这 7 人 中包括 来自 法国 和 俄罗斯 的 宇航 员．
这7人中包括来自法国和俄罗斯的宇航员。
这些 7 人 中包括 来自 法国 和 俄罗斯 的 宇航 员 。
这7人中包括来自法国和俄罗斯的宇航员。
These 7 people include astronauts coming from France and Russia.
These 7 people include astronauts coming from France and Russia.
“Here, dogs are wearing the mouth, and have masters traction.”
Syntax-Based Summary

• Lots of technology...
• Critical part: underlying “generative story”

How does a tree in language X become a string in language Y?

“... through a sequence of steps of some tree automaton operating on Penn Treebank-style trees...”
Morphology?

• Syntax-based MT systems currently work at the word level
• Possible direction:
  – Morpho-syntactic translation models

How does a **morpho-syntactic tree** in language X become a **string of characters** in language Y?
Current Syntax-Based SMT

RULE BASE

q.JJ(red) <-> rojo q.JJ(green) <-> verde
q.JJ(red) <-> roja q.JJ(green) <-> verdes
q.JJ(red) <-> rojos q.N(cat) <-> gato
q.JJ(red) <-> rojas q.N(cats) <-> gatos
q.N(car) <-> coche q.N(moon) <-> luna
q.N(cars) <-> coches q.N(moons) <-> lunas
q.DT(a) <-> un q.N(light) <-> luz
q.DT(a) <-> una q.N(lights) <-> luzes

\( q.NP(x0:DT \ x1:JJ \ x2:N) \) <-> q.x0 q.x2 q.x1

Very large
wordform-to-wordform
dictionary

Simple syntactic
combination
Current Syntax-Based SMT

Original input:

Transformation:
Current Syntax-Based SMT

Original input:

Transformation:
Current Syntax-Based SMT

Original input: a red car

Transformation: q DT, q NN, q JJ
a car red
Current Syntax-Based SMT

Original input:

```
NP
   /\ 
  DT  JJ  N
   a   red  car
```

Transformation:

```
q DT \ 0.2
   \   una
     a
```

```
q DT, q NN, q JJ
   a, car, red
```
Current Syntax-Based SMT

Original input: a red car

Transformation: una, q NN, q JJ, car, red
Current Syntax-Based SMT

Original input:

Transformation:
Current Syntax-Based SMT

Original input:

Transformation:
Current Syntax-Based SMT

Original input:

Transformation:

una, coche, rojas
Current Syntax-Based SMT

% echo 'NP(DT(a) JJ(red) N(car))' |
tiburon -l -k 8 - sbmt.xlnts

OUTPUTS:

una coche rojo # 1.0  un coche rojas # 1.0
una coche rojas # 1.0  una coche roja # 1.0
un coche rojos # 1.0   un coche rojo # 1.0
una coche rojos # 1.0  un coche roja # 1.0

Overgeneration
(rely on language model to catch problems)
Possible Morpho-Syntactic SMT?

RULE BASE

qjo.red <-> r o j
qnsmasc.car <-> c o c h e
qnsfem.moon <-> l u n a
qdmasc.a <-> u n
qje.green <-> v e r d e
qnsmasc.cat <-> g a t o
qnesfem.light <-> l u z
qdfem.a <-> u n a

qmasc.JJ(x0:) <-> qjo.x0 o
qmasc.JJ(x0:) <-> qje.x0
qplmasc.x0:JJ <-> qmasc.x0 s
qplmasc.N(x0: x1:pl) <-> qnsmasc.x0 s
qplmasc.N(x0: x1:pl) <-> qnesmasc.x0 e s

...

q.NP(x0:DT x1:JJ x2:N) <-> qdmasc.x0 _ qmasc.x2 _ qmasc.x1

% echo 'NP(DT(a) JJ(red) N(car))' | tiburon -l -k 1 - msmt.xlnts

OUTPUTS:

u n _ c o c h e _ r o j o # 1.0 (no other outputs)
Possible Morpho-Syntactic SMT?

Original input:

Transformation:
Possible Morpho-Syntactic SMT?

Original input: a red car

Transformation:
q NP  0.2  qdmasc x0 , _ , qmasc x2 , _ , qmasc x1
x0:DT  x1:JJ  x2:N

q NP
DT  JJ  N
a  red  car
Possible Morpho-Syntactic SMT?

Original input:  

Transformation:  

```
q NP  \rightarrow qdmasc x0,_, qmasc x2,_, qmasc x1
x0:DT x1:JJ x2:N
```

```
qdmasc DT ,_, qmasc N ,_, qmasc JJ
a car red
```
Possible Morpho-Syntactic SMT?

Original input:  
Transformation:
Possible Morpho-Syntactic SMT?

Original input:

Transformation:
Possible Morpho-Syntactic SMT?

Original input:

Transformation:
Possible Morpho-Syntactic SMT?

Original input:

Transformation:
Possible Morpho-Syntactic SMT?

Original input: 

Transformation: 

```
NP
  /   
DT JJ N
  a  red car

qjo red \rightarrow r, o, j

u, n, _, qmasc N, _ qjo red, o
car
```
Possible Morpho-Syntactic SMT?

Original input:

Transformation:

Original input: a red
transformed to: r, o, j

Transformation: u, n, _, qmasc N, _, r, o, j, o
Possible Morpho-Syntactic SMT?

Original input: a red car

Transformation:

u, n _, c, o, c, h e, _, r, o, j, o
Possible Morpho-Syntactic SMT?

% echo 'NP(DT(a) JJ(red) N(car))' |
tiburon -l -k 1 - msmt.xlnts

OUTPUTS:

un_coches_rojo # 1.0 (no other outputs)
Another Open Problem

• Syntax-based MT systems acquire rules from aligned, parsed parallel text
• But alignments are still done without syntax, via GIZA++ and are poor
  – Brown et al 1990 technology!
Another Open Problem

- Syntax-based MT systems acquire rules from aligned, parsed parallel text
- But alignments are still done without syntax, via GIZA++ and are poor
  - Brown et al 1990 technology!
Another Open Problem

• Syntax-based MT systems acquire rules from aligned, parsed parallel text
• But alignments are still done without syntax, via GIZA++ and are poor
  — Brown et al 1990 technology!

Use generalized morpho-syntactic framework to explain (align) parallel data at character level.
Like GIZA++, do it unsupervised.
Summary

• Some ancient history

• Some of what’s happening in morphology and MT

• Some possible directions
thanks