Paraphrasing

May 1, 2012

Goals of today's lecture

- Understand what paraphrases are
- Discuss how we can can re-use MT machinery of for other text-to-text (T2T) generation tasks
- Review various data-driven methods for learning paraphrases
- Focus on a method that uses bilingual pivoting
- Define a set of modifications that we need to make to the MT pipeline to customize it to new tasks

What are Paraphrases?

Differing textual expressions of the **same meaning**:

cup	mug
the king's speech	His Majesty's address
X_1 talks to X_2	X_1 converses with X_2
NN devoured NP	NP was eaten by NN
Many Republicans' hearts were broken by Chris Christie reiterating his refusal to run for the presidency.	The Garden State governor stated once again that he will not seek the presidential nomination, disappointing Republicans.

What are they good for?

Anything that deals with **text** and **meaning**, i.e. automatic...

...summarization, translation, MT evaluation, question answering, information retrieval, natural language generation, essay grading, sentiment analysis, linguistic stenography, entailment recognition, etc.

Real question is **where do we get them**?

Many NLP tasks can be viewed as "MT"

- If you have a "source" and a "target" that are aligned on the sentence-level, then you can re-use much of the MT machinery to "translate" between them
- Input this parallel corpus and then re-use
  - Word alignment algorithms
  - Phrase table extraction
  - Decoder + LM
- Example task: Sentence simplification

Regular English-Simple English Parallel Corpus

a synonym for " lolcat " is cat macro , since the images are a type of image macro .	a different word for lolcat is cat macro because it is a kind of image macro .
genetic engineering has expanded the genes available to breeders to utilize in creating desired germlines for new crops .	new plants were created with genetic engineering .
the dominant classical dance amongst tamils is bharatanatyam .	bharatanatyam is the main dance of the tamil people .
a naval mine is a self-contained explosive device placed in water to destroy ships or submarines .	a naval mine is a bomb placed in water to destroy ships or submarines .

Word align the parallel corpus

a **synonym** for " lolcat " is cat macro **, since** the images **are** a **type** of image macro

a **different word** for lolcat is cat macro **because** it **is** a **kind** of image macro .

Extract phrase table

synonym	|	different word
, since	|	because
are	|	is
type	|	kind
a synonym for " X " is Y	|	a different word for X is Y

Decode

since then they have changed their name to palladium and played **alongside** amy winehouse .

phrase table +
simple English LM

since then **,** they have changed their name to palladium and played **with** amy winehouse.

Done! Right??

- Just need to calculate a BLEU score and then write a paper

- What is wrong with this?
- Where does it get things right and where does it get things wrong?
- (Discuss with your neighbor)

Paraphrasing with parallel monolingual data

- Some work has use parallel monolingual data
- Comparable corpora
  - Encyclopedia articles on same topic
  - Different newspapers' accounts of one event
- Multiple translations of the same foreign text
  - Evaluation data for Bleu metric
  - Different translations of classic French novels into English

What a scene! Seized by the tentacle and **glued to** its suckers, the unfortunate man was **swinging in the air** at the **mercy** of this enormous appendage. He gasped, he choked, he yelled: "Help! Help!" I'll hear his **harrowing plea** the rest of my life!
The **poor fellow** was **done for**.

What a scene! The unhappy man, seized by the tentacle and **fixed to** its suckers, was **balanced in the air** at the **caprice** of this enormous trunk. He rattled in his throat, he was stifled, he cried, "Help! help!" That **heart-rending cry**! I shall hear it all my life.
The **unfortunate man** was **lost**.

Paraphrasing with parallel monolingual data

- Barzilay and McKeown (2001) used identical contexts in aligned sentences:

Emma burst into tears **and he tried to** comfort **her, saying things to make her smile.**
Emma cried **and he tried to** console **her, adorning his words with puns.**

- burst into tears = cried and comfort = console

Potential problems with these methods

- Multiple translations are relatively uncommon
- This Limits what paraphrases we can generate
  - Limited number of paraphrases
  - Constrained to a few genres

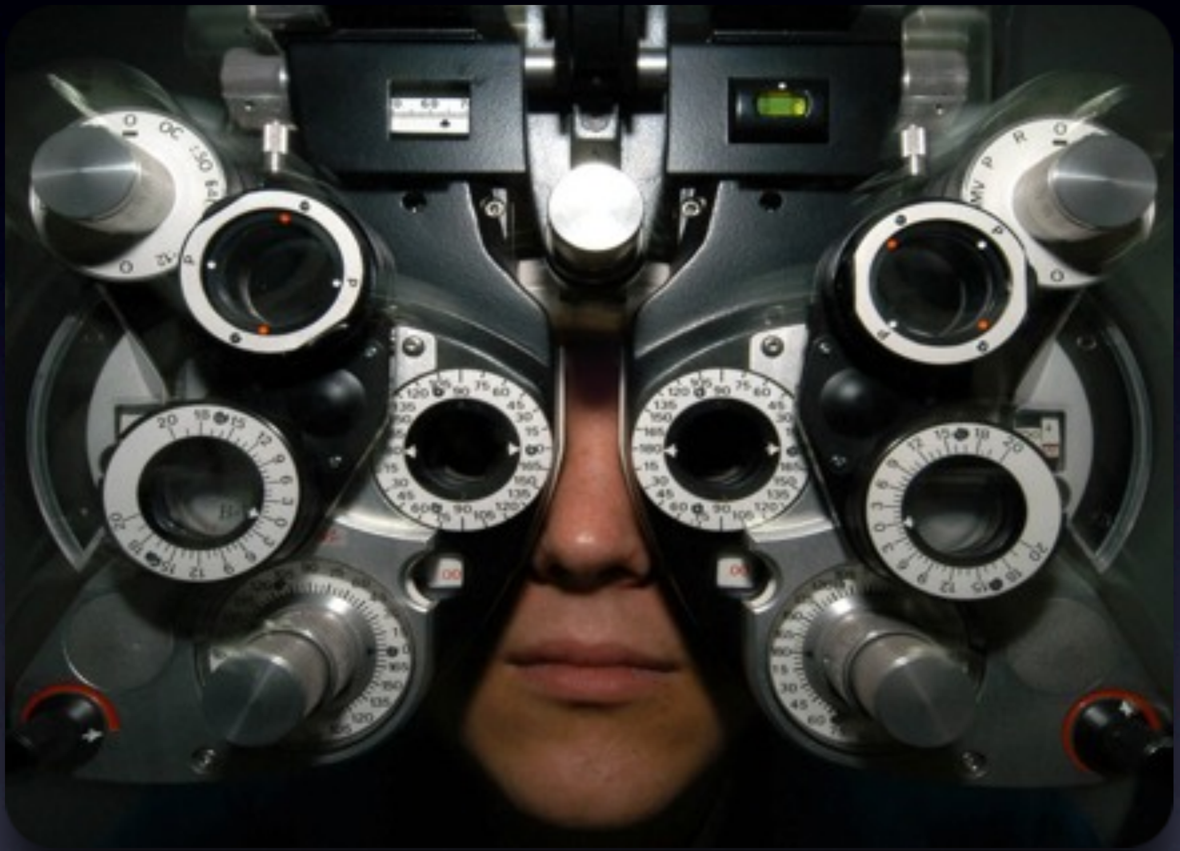
Distributional Hypothesis

If we consider oculist and eye-doctor we find that, as our corpus of utterances grows, these two occur in almost the same environments. In contrast, there are many sentence environments in which oculist occurs but lawyer does not...

It is a question of the relative frequency of such environments, and of what we will obtain if we ask an informant to substitute any word he wishes for oculist (not asking what words have the same meaning).

These and similar tests all measure the probability of particular environments occurring with particular elements... If A and B have almost identical environments we say that they are synonyms.

–Zellig Harris (1954)



Duty and Responsibility

- To operationalize the Distributional Hypothesis we must define similar environments
- Lin and Panel (2001) used dependency relationships
- Duty and responsibility share a similar set of dependency contexts in large volumes of text:

modified by adjectives	objects of verbs
additional, administrative, assigned, assumed, collective, congressional, constitutional ...	assert, assign, assume, attend to, avoid, become, breach ...

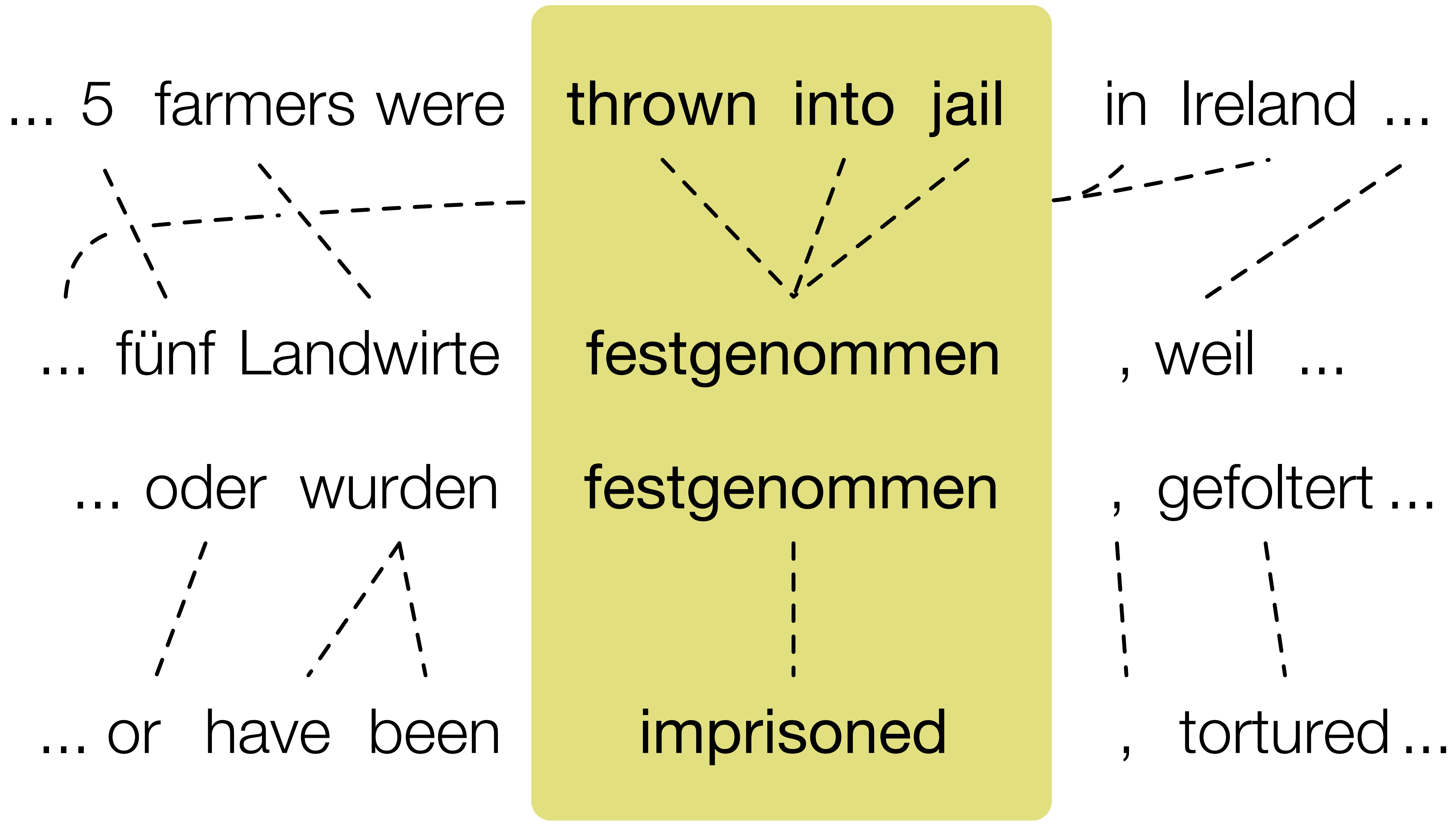
Problem with distributional similarity

- Distributional methods group related words that are not synonymous:
  - cats and dogs, girls and boys

Paraphrasing with Bilingual parallel corpora

- Bilingual parallel corpora are much more common than monolingual parallel corpora
- However, no longer contain identical contexts
- Use aligned foreign language phrase as pivot
- Less prone to retrieve non-synonymous related words

Bilingual pivoting



the establishment of the military force is in their view a tool to realise these aims
die bildung einer truppe ist ihrer auffassung nach ein mittel zur durchsetzung dieser ziele
es ist eine truppe die aus nationalen einheiten besteht
it will be a force comprised of various national units
the 1000 strong military force will be involved in peacemaking
die 1000 mann starke friedenstruppe soll zur friedensschaffung herangezogen werden
hat entführungen der uno friedenstruppe verurteilt
condemned the abductions of un peace-keeping personnel
the eu may carry out tasks which do not use military force
die eu sollte aufgaben durchführen bei denen keine streitkräfte zum einsatz kommen
angola beispielsweise besitzt starke streitkräfte die wertvolle hilfe hätten leisten können
angola for example has powerful armed forces which could have given valuable assistance
the eu may carry out tasks which do not use military force
die eu sollte aufgaben durchführen bei denen keine streitkräfte zum einsatz kommen
aufgrund eines gekürzten verteidigungshaushaltes können die streitkräfte gegenwärtig jedoch nur etwa 20,000 mann aufbringen
due to reduced defence spending the national defense can currently only supply approximately 20,000 men

Many, many alternatives

Paraphrase candidates for “thrown into jail”

Good
jailed
arrested
imprisoned
incarcerated
locked up
taken into custody
thrown into prison

Bad / Ugly
being arrested
in jail
put in prison for
maltreated
thrown
cases
custody

Good examples

- dead bodies → corpses, carcasses, bodies, skeletons, people
- military force → force, forces, peace-keeping personnel, armed forces, military forces, defense
- sooner or later → eventually, at some point
- wish to clarify → want to make perfectly clear, would like to ask, would like to comment on, would like to mention, would like to deal with, would comment on
- every other → any other, all, other, every, all other, everyone else, others, all the others

Bad examples

- are perfectly entitled → perfectly entitled, have every right, right, are, has a legitimate, call for, has, legitimate right, have the right
- for small-scale projects → small-scale projects, small, of, only trifling amounts are at stake, for projects, for smaller-scale projects, to, for smaller projects
- groundwork for → for, groundwork, to, basis for, the, basis, preparation, foundations for, that
- create equal → equal, to create a, create, to create equality, same, created, conditions

Separating the Good from the Bad

- How could we differentiate good paraphrases from bad ones?

- (Discuss with your neighbor)

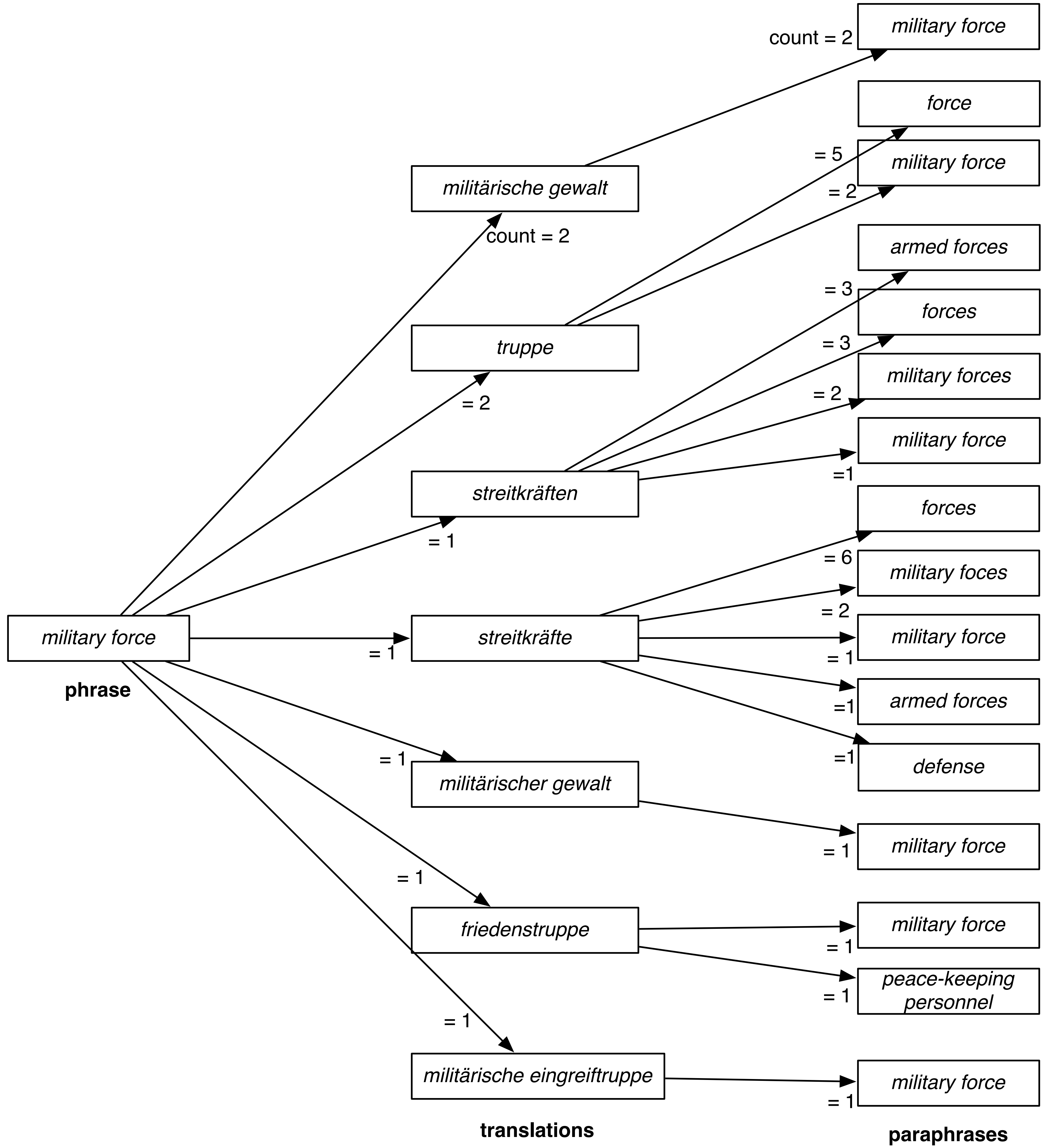
Ranking Paraphrases

- Paraphrase probability

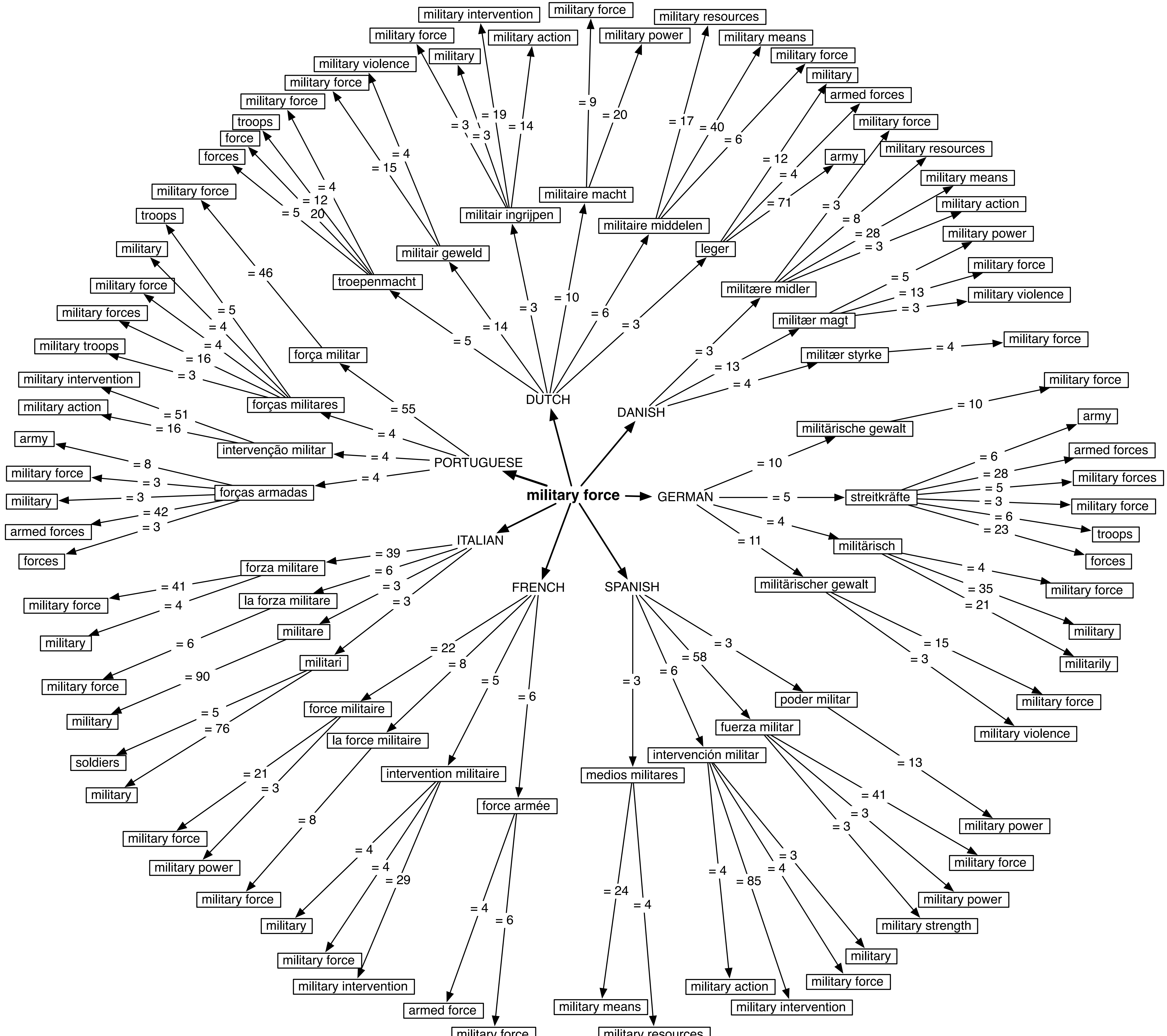
$$
\begin{aligned}
p(e_2|e_1) &= \sum_f p(e_2, f|e_1) \\
&= \sum_f p(e_2|f, e_1)p(f|e_1) \\
&\approx \sum_f p(e_2|f)p(f|e_1)
\end{aligned}
$$

$$
p(f|e) = \frac{count(e, f)}{\sum_f count(e, f)}
$$

Log-linear model with additional features

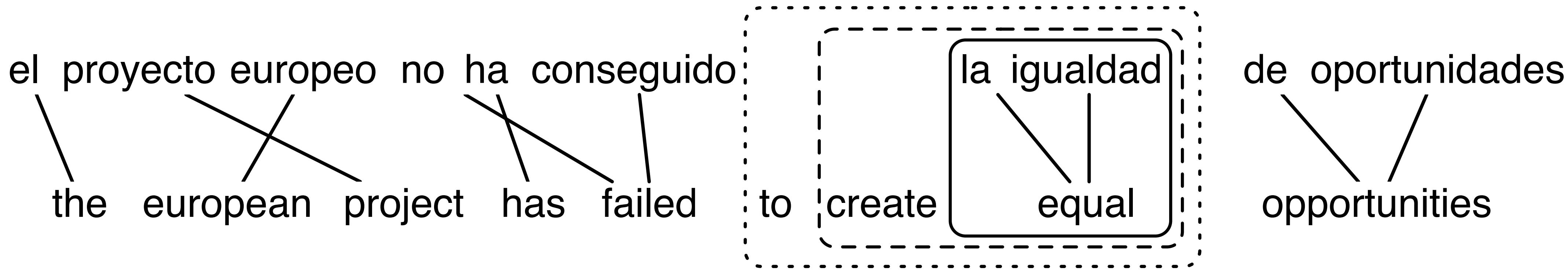


military force
phrase
militärische gewalt
count = 2
truppe
= 2
streitkräften
= 1
streitkräfte
= 1
militärischer gewalt
= 1
friedenstruppe
= 1
militärische eingreiftruppe
= 1
translations
military force
count = 2
force
= 5
military force
= 2
armed forces
= 3
forces
= 3
military forces
= 2
military force
=1
forces
= 6
military foces
= 2
military force
= 1
armed forces
=1
defense
=1
military force
= 1
military force
= 1
peace-keeping personnel
= 1
military force
= 1
paraphrases



military force
DUTCH
DANISH
PORTUGUESE
GERMAN
ITALIAN
FRENCH
SPANISH
troepenmacht
militair geweld
militair ingrijpen
militaire macht
militaire middelen
leger
= 5
= 14
= 3
= 10
= 6
= 3
forces
force
troops
military force
military force
military violence
= 5
= 20
= 12
= 4
= 15
= 4
military
military force
military intervention
military action
= 3
= 3
= 19
= 14
military force
military power
= 9
= 20
military resources
military means
military force
= 17
= 40
= 6
military
armed forces
army
= 12
= 4
= 71
militære midler
militær magt
militær styrke
= 3
= 13
= 4
military force
military resources
military means
military action
= 3
= 8
= 28
= 3
military power
military force
military violence
= 5
= 13
= 3
military force
= 4
força militar
forças militares
intervenção militar
forças armadas
= 55
= 4
= 4
= 4
military force
= 46
troops
military
military force
military forces
military troops
= 5
= 4
= 4
= 16
= 3
military intervention
military action
= 51
= 16
army
military force
military
armed forces
forces
= 8
= 3
= 3
= 42
= 3
militärische gewalt
streitkräfte
militärisch
militärischer gewalt
= 10
= 5
= 4
= 11
military force
= 10
army
armed forces
military forces
military force
troops
forces
= 6
= 28
= 5
= 3
= 6
= 23
military force
military
militarily
= 4
= 35
= 21
military force
military violence
= 15
= 3
forza militare
la forza militare
militare
militari
= 39
= 6
= 3
= 3
military force
military
= 41
= 4
military force
= 6
military
= 90
soldiers
military
= 5
= 76
force militaire
la force militaire
intervention militaire
force armée
= 22
= 8
= 5
= 6
military force
military power
= 21
= 3
military force
= 8
military
military force
military intervention
= 4
= 4
= 29
armed force
military force
= 4
= 6
medios militares
intervención militar
fuerza militar
poder militar
= 3
= 6
= 58
= 3
military means
military resources
= 24
= 4
military action
military intervention
military force
military
= 4
= 85
= 4
= 3
military force
military power
military strength
= 41
= 3
= 3
military power
= 13

Phrase extraction with unaligned words



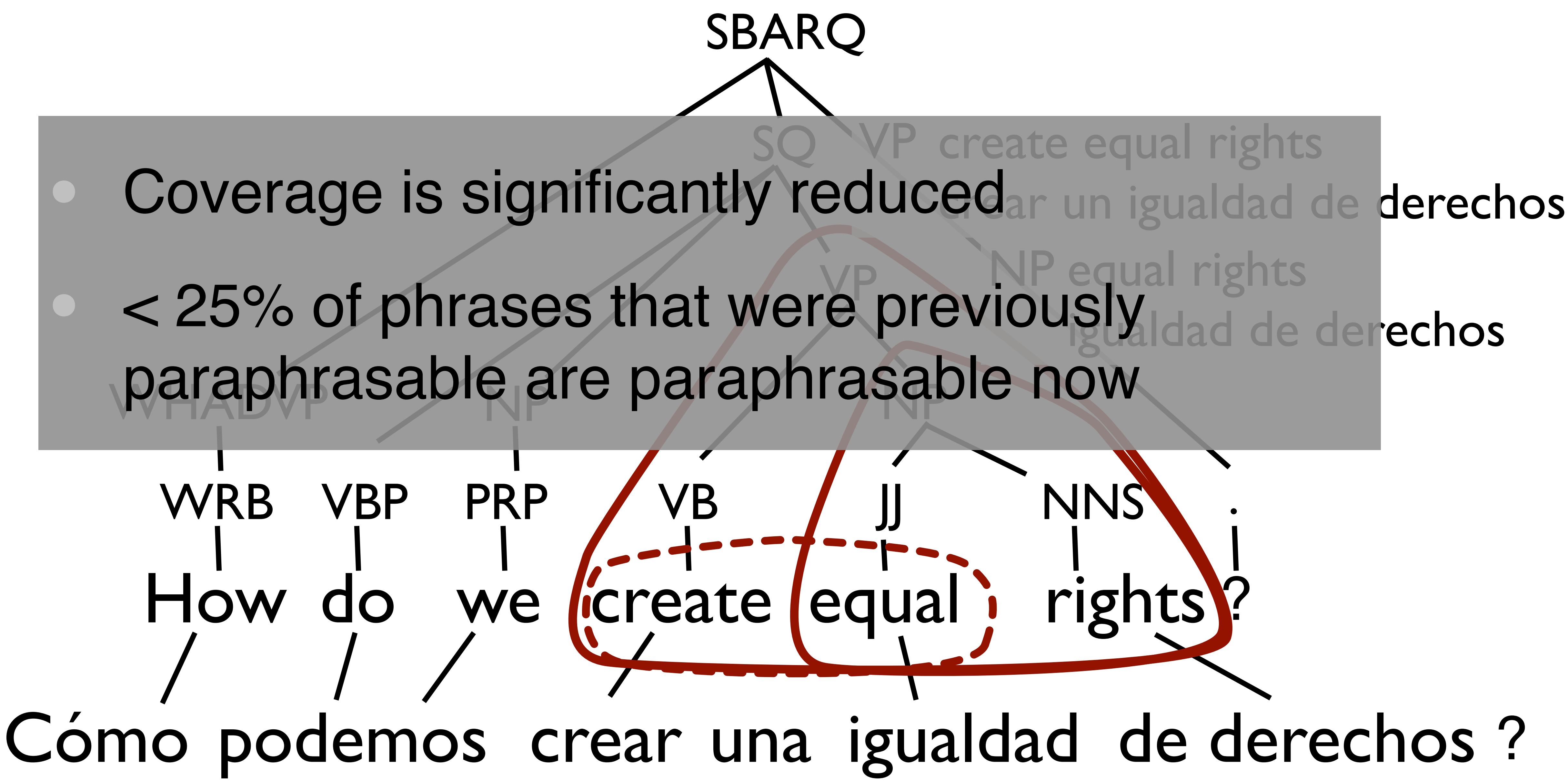
la igualdad = equal create equal to create equal

- For 3.7m paraphrases of 400k phrases
  - 34% were sub- or super-strings
  - 73% of the paraphrases that were ranked highest by the paraphrase probability

Syntactic Constraints

- Require phrases and their paraphrase to be the same syntactic type
- Redefine the paraphrase probability to condition on syntactic labels
- Change the phrase extraction algorithm so that it enumerates phrase pairs and syntactic labels

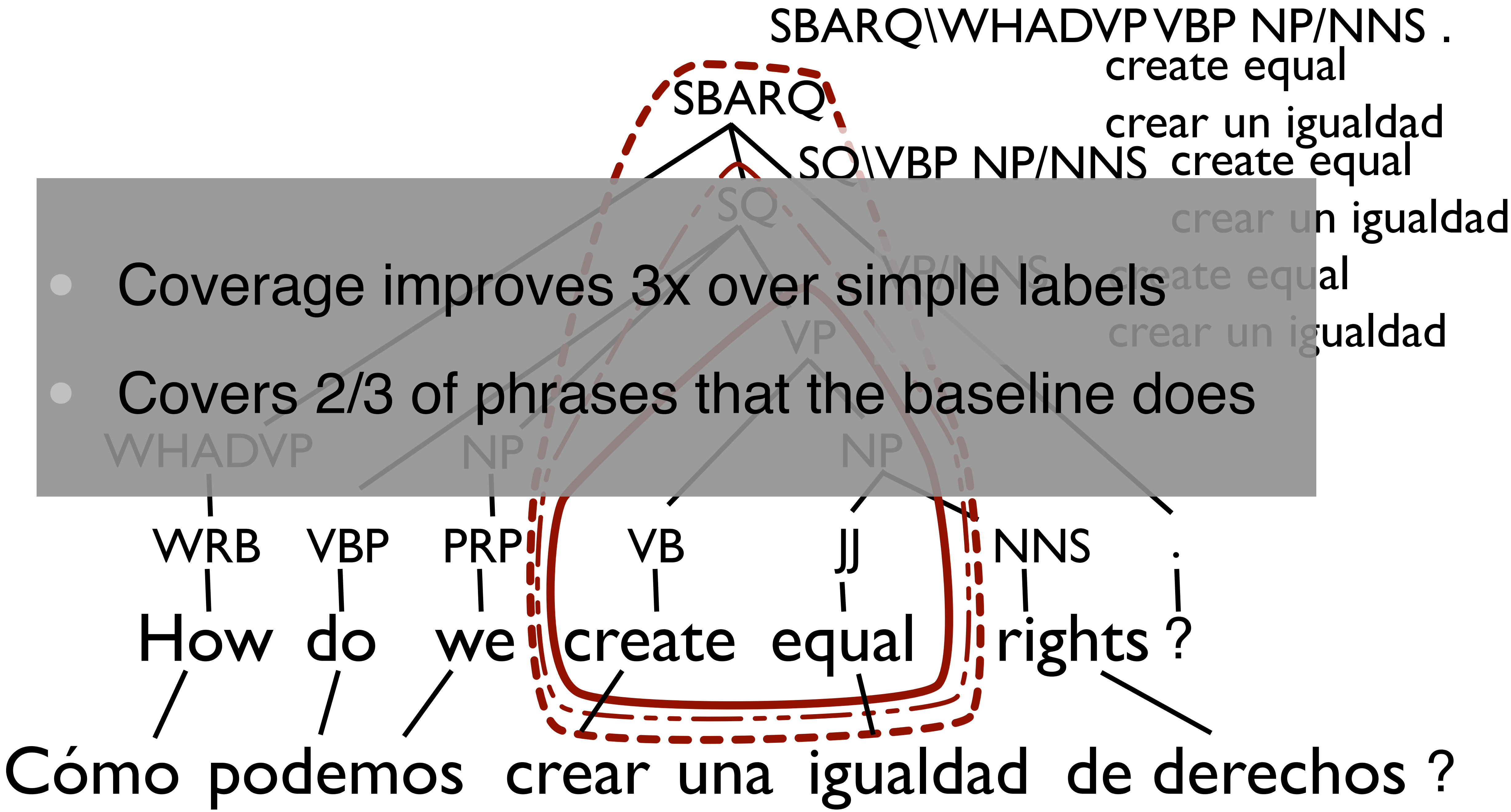
Phrase extraction + syntactic labels



- Coverage is significantly reduced
- < 25% of phrases that were previously paraphrasable are paraphrasable now

Using complex labels

- Coverage improves 3x over simple labels
- Covers 2/3 of phrases that the baseline does



Example improvements

- create equal | equal, to create a, create, to create equality, same, created, conditions

- VP/NNS → create equal | creating equal
- VP/NNS PP → create equal | promote equal, establish fair
- VP/NNS PP PP → create equal | creating equal, provide equal, create genuinely fair

Example improvements

- equal | same, equality, equals, equally, the, fair, equal rights
- JJ → equal | same, fair, similar, equivalent
- ADJP → equal | necessary, similar, identical, the same, equal in law, equivalent

SCFGs for Paraphrasing

- What does this notation remind you of?
  - JJ → equal | same

- Synchronous context free grammars!
- If you hadn't guessed already, we can fuse the idea of pivoting with syntactic MT to get SCFGs for paraphrasing

Meaning preserving transformations

- Adapting our syntactic MT models, we learn structural transformations, like the English possessive rule

$$NP \rightarrow \quad NP\ \text{'s}\ NN \mid \text{le}\ NN\ \text{de}\ NP$$

$$NP \rightarrow \quad \text{the}\ NN\ \text{of}\ NP \mid \text{le}\ NN\ \text{de}\ NP$$

combine to

$$NP \rightarrow \quad NP\ \text{'s}\ NN \mid \text{the}\ NN\ \text{of}\ NP$$

Possessive rule	NP $\rightarrow$	the NN of the NNP	the NNP's NN
	NP $\rightarrow$	the NNS_1 made by the NNS_2	the NNS_2's NNS_1
Dative shift	VP $\rightarrow$	give NN to NP	give NP the NN
	VP $\rightarrow$	provide NP_1 to NP_2	give NP_2 NP_1
Adv./adj. phrase move	S/VP $\rightarrow$	ADVP they VBP	they VBP ADVP
	S $\rightarrow$	it is ADJP VP	VP is ADJP
Verb particle shift	VP $\rightarrow$	VB NP up	VB up NP
Reduced relative clause	SBAR/S $\rightarrow$	although PRP VBP that	although PRP VBP
	ADJP $\rightarrow$	very JJ that S	JJ S
Partitive constructions	NP $\rightarrow$	CD of the NN	CD NN
	NP $\rightarrow$	all DT\NP	all of the DT\NP
Topicalization	S $\rightarrow$	NP, VP.	VP, NP.
Passivization	SBAR $\rightarrow$	that NP had VBN	which was VBN by NP
Light verbs	VP $\rightarrow$	take action ADVP	to act ADVP
	VP $\rightarrow$	to take a decision PP	to decide PP

Sentential Paraphrasing

- These paraphrasing SCFGs can be used for monolingual text-to-text generation tasks
- Non-naive reuse of SMT machinery
- Adapt translation framework with appropriate
  - Development data
  - Objective function
  - Feature sets
  - Grammar augmentations

Example: Sentence Compression

- Problem: given an input sentence, rewrite it into a shorter sentence while preserving the core meaning:

and he said that the project will cover the needs of the region in the long term.

he said the project includes all the district's long-term needs.

SMT Machinery

- What we can directly re-use:
  - Grammar extraction & formalism
  - Decoding & n-gram language model integration
  - Log-linear model formulation
  - MERT for parameter tuning

SMT Machinery

Development Data	Multi-reference sets
Objective Function	BLEU
Features	$P_{phrase}(e_1 \mid e_2)$, $P_{lex}(e_1 \mid e_2)$

Reusing SMT for Text-to-Text

- Inter-reference BLEU is typically very high (52.7)
- Resulting paraphrases are almost always identity

Input	the election campaign , which did not gain the interest of voters , ended friday .
Paraphrase	the election campaign , which did not gain the interest of voters , ended friday .

Adapting SMT Machinery

	SMT	Sentence Compression
Development Data	Multi-reference sets	<sentence, compression>
Objective Function	BLEU	cmpBLEU
Features	$P_{phrase}(e_1 \mid e_2)$, $P_{lex}(e_1 \mid e_2)$	+ length(e_1), length(e_2), length_diff(e_1, e_2), etc
Augmentation	n/a	Deletion rules

Development Data

- Common compression corpora are deletion-based (e.g. Ziff-Davis)
- We create a development and test sets from reference translations for SMT
- Consists of compressive sentential paraphrases (CR 0.8 to 0.5, 0.73 avg.)

and he said that the project will cover the needs of the region in the long term.

he said the project includes all the district's long-term needs.

Objective Function

- Penalize insufficient compressions
- Reward well-formed language
- Penalize overzealous compressions

$$\textsc{cmp}\textsc{Bleu}_{\lambda,\theta}(i, o) = \begin{cases} e^{\lambda(\theta - c)} \cdot \textsc{Bleu}(o) & \text{if } c > \theta \\ \textsc{Bleu}(o) & \text{otherwise} \end{cases}$$

Feature Functions

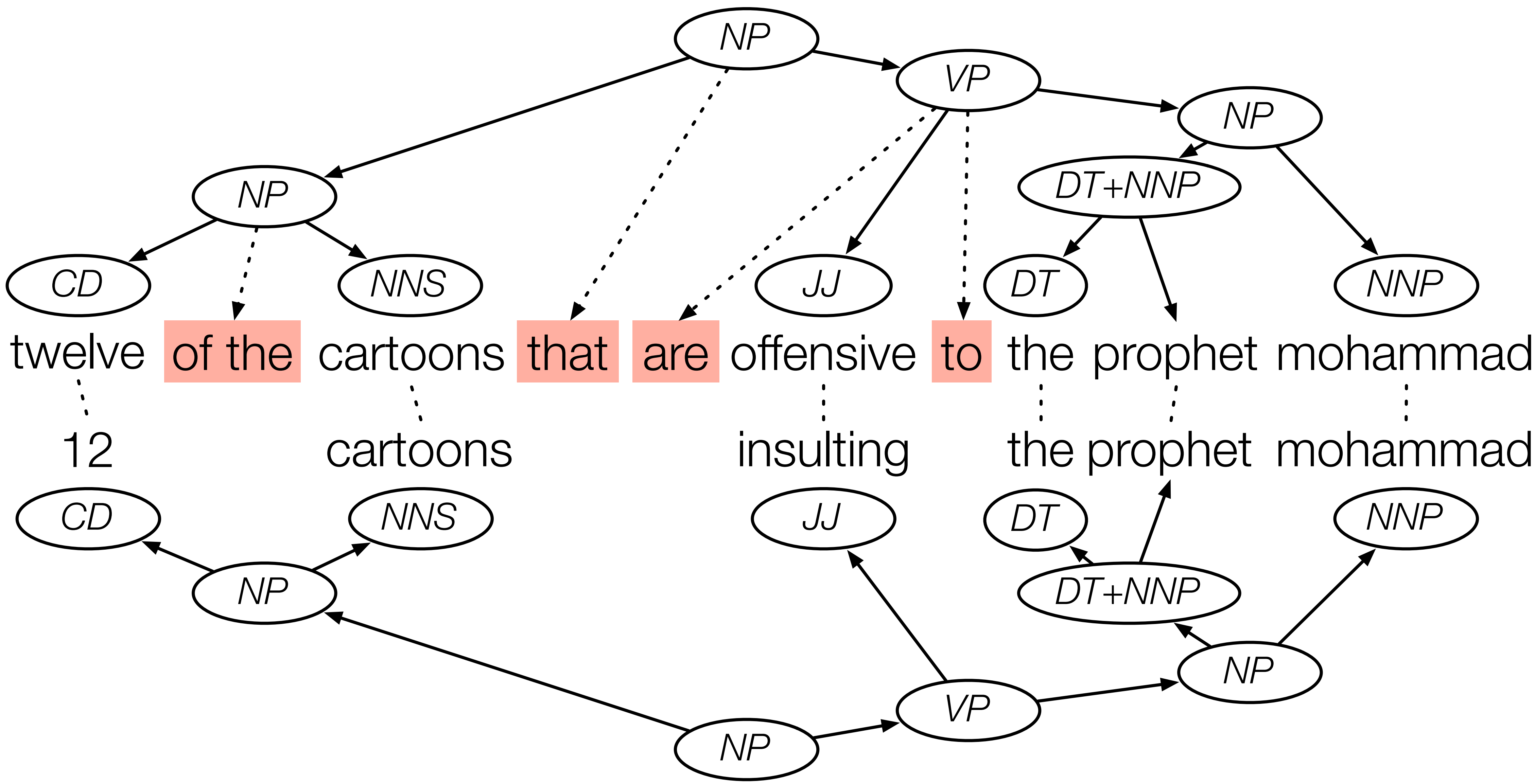
- Augment rules with length information
  - –Number of words on source & target side
  - –Difference in number of words
  - –Difference in number of characters

Grammar Augmentations

- Added deletion rules for hand-chosen POS
  - JJ, JJR, JJS
  - RB, RBR, RBS
  - DT

$$JJ \rightarrow \quad \text{superfluous} \quad | \quad \varepsilon$$

Example sentence compression



Lexical paraphrase:

JJ → offensive | insulting

Reduced relative clause:

NP → NP that VP | NP VP

Pred. adjective copula deletion:

VP → are JJ to NP | JJ NP

Partitive construction:

NP → CD of the NNS | CD NNS

Text-to-text generation tasks

- Sentence compression
- Sentence simplification
- English as a Second Language (ESL) error correction
- Poetry generation
- Legalese to plain English translation

Conclusions

- Paraphrases are useful for a wide range of NLP tasks
- Tempting to think of SMT as a tool that can be used to do anything ... just find “parallel corpus”
- Doesn’t work well if done simplemindedly
- Better to extract paraphrases from bilingual parallel corpora
- Then adapt the SMT machinery in non-naive ways