6.864 (Fall 2007)

Machine Translation Part IV

Overview

- Syntax Based Model 1: (Wu 1995)

(Wu 1995)

- Standard probabilistic context-free grammars:
  probabilities over rewrite rules define probabilities over trees, strings, in one language

- **Transduction grammars**:
  Simultaneously generate strings in two languages

A Probabilistic Context-Free Grammar

S	⇒	NP	VP	1.0
VP	⇒	Vi		0.4
VP	⇒	Vt	NP	0.4
VP	⇒	VP	PP	0.2
NP	⇒	DT	NN	0.3
NP	⇒	NP	PP	0.7
PP	⇒	P	NP	1.0

Vi	⇒	sleeps	1.0
Vt	⇒	saw	1.0
NN	⇒	man	0.7
NN	⇒	woman	0.2
NN	⇒	telescope	0.1
DT	⇒	the	1.0
IN	⇒	with	0.5
IN	⇒	in	0.5

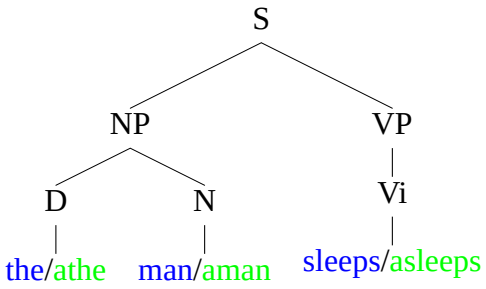
- Probability of a tree with rules $\alpha_i \rightarrow \beta_i$ is $\prod_i P(\alpha_i \rightarrow \beta_i | \alpha_i)$

Transduction PCFGs

- First change to the rules: **lexical** rules generate a pair of words

Vi	$\Rightarrow$	sleeps/asleeps	1.0
Vt	$\Rightarrow$	saw/asaw	1.0
NN	$\Rightarrow$	man/aman	0.7
NN	$\Rightarrow$	woman/awoman	0.2
NN	$\Rightarrow$	telescope/atelescope	0.1
DT	$\Rightarrow$	the/athe	1.0
IN	$\Rightarrow$	with/awith	0.5
IN	$\Rightarrow$	in/ain	0.5

Transduction PCFGs



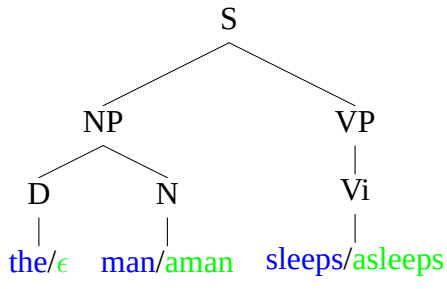
- The modified PCFG gives a distribution over $(\mathbf{f}, \mathbf{e}, T)$ triples, where $\mathbf{e}$ is an English string, $\mathbf{f}$ is a French string, and T is a tree

Transduction PCFGs

- Another change: allow empty string ϵ to be generated in either language, e.g.,

DT	$\Rightarrow$	the/ϵ	1.0
IN	$\Rightarrow$	ϵ/awith	0.5

Transduction PCFGs



- Allows strings in the two languages to have different lengths

  the man sleeps $\Rightarrow$ aman asleeps

Transduction PCFGs

- Final change: currently formalism does not allow different word orders in the two languages
- Modify the method to allow two types of rules, for example

$$S \Rightarrow [NP\ VP] \quad 0.7$$

$$S \Rightarrow \langle NP\ VP \rangle \quad 0.3$$

- Defi ne:
  - E_X is the English string under non-terminal X
    e.g., E_{NP} is the English string under the NP
  - F_X is the French string under non-terminal X
- Then for S $\Rightarrow$ [NP VP] we defi ne

$$E_S = E_{NP}.E_{VP}$$
$$F_S = F_{NP}.F_{VP}$$

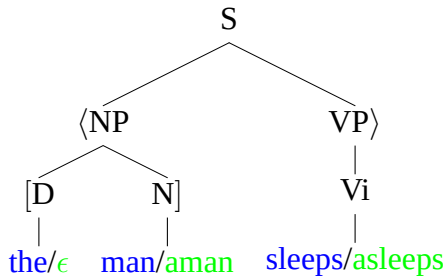
where . is concatentation operation

- For S $\Rightarrow \langle$NP VP$\rangle$ we defi ne

$$E_S = E_{NP}.E_{VP}$$
$$F_S = F_{VP}.F_{NP}$$

In the second case, the string order in French is reversed

Transduction PCFGs



- This tree represents the correspondance

the man sleeps ⇒ asleeps aman

A Transduction PCFG

S	⇒	[NP VP]	0.7
S	⇒	⟨NP VP⟩	0.3
VP	⇒	Vi	0.4
VP	⇒	[Vt NP]	0.01
VP	⇒	⟨Vt NP⟩	0.79
VP	⇒	[VP PP]	0.2
NP	⇒	[DT NN]	0.55
NP	⇒	⟨DT NN⟩	0.15
NP	⇒	[NP PP]	0.7
PP	⇒	⟨P NP⟩	1.0

Vi	$\Rightarrow$	sleeps/ϵ	0.4
Vi	$\Rightarrow$	sleeps/asleeps	0.6
Vt	$\Rightarrow$	saw/asaw	1.0
NN	$\Rightarrow$	ϵ/aman	0.7
NN	$\Rightarrow$	woman/awoman	0.2
NN	$\Rightarrow$	telescope/atelescope	0.1
DT	$\Rightarrow$	the/athe	1.0
IN	$\Rightarrow$	with/awith	0.5
IN	$\Rightarrow$	in/ain	0.5

(Wu 1995)

- Dynamic programming algorithms exist for "parsing" a pair of English/French strings (finding most likely tree underlying an English/French pair). Runs in $O(|\mathbf{e}|^3|\mathbf{f}|^3)$ time.

- Training the model: given $(\mathbf{e}_k, \mathbf{f}_k)$ pairs in training data, the model gives

$$P(T, \mathbf{e}_k, \mathbf{f}_k \mid \Theta)$$

where T is a tree, Θ are the parameters. Also gives

$$P(\mathbf{e}_k, \mathbf{f}_k \mid \Theta) = \sum_T P(T, \mathbf{e}_k, \mathbf{f}_k \mid \Theta)$$

Likelihood function is then

$$L(\Theta) = \sum_k \log P(f_k, e_k \mid \Theta) = \sum_k \log \sum_T P(T, f_k, e_k \mid \Theta)$$

Wu gives a dynamic programming implementation for EM

R:	the current diffi culties should encourage us to redouble our efforts to promote cooperation in the euro-mediterranean framework.
C:	the current problems should spur us to intensify our efforts to promote cooperation within the framework of the europa-mittelmeerprozesses.
B:	the current problems should spur us, our efforts to promote cooperation within the framework of the europa-mittelmeerprozesses to be intensifi ed.
R:	propaganda of any sort will not get us anywhere.
C:	with any propaganda to lead to nothing.
B:	with any of the propaganda is nothing to do here.
R:	yet we would point out again that it is absolutely vital to guarantee independent fi nancial control.
C:	however, we would like once again refer to the absolute need for the independence of the fi nancial control.
B:	however, we would like to once again to the absolute need for the independence of the fi nancial control out.
R:	i cannot go along with the aims mr brok hopes to achieve via his report.
C:	i cannot agree with the intentions of mr brok in his report persecuted.
B:	i can intentions, mr brok in his report is not agree with.
R:	on method, i think the nice perspectives, from that point of view, are very interesting.
C:	what the method is concerned, i believe that the prospects of nice are on this point very interesting.
B:	what the method, i believe that the prospects of nice in this very interesting point.

R:	secondly, without these guarantees, the fall in consumption will impact negatively upon the entire industry.
C:	and, secondly, the collapse of consumption without these guarantees will have a negative impact on the whole sector.
B:	and secondly, the collapse of the consumption of these guarantees without a negative impact on the whole sector.
R:	awarding a diploma in this way does not contravene uk legislation and can thus be deemed legal.
C:	since the award of a diploms is not in this form contrary to the legislation of the united kingdom, it can be recognised as legitimate.
B:	since the award of a diploms in this form not contrary to the legislation of the united kingdom is, it can be recognised as legitimate.
R:	i should like to comment briefly on the directive concerning undesirable substances in products and animal nutrition.
C:	i would now like to comment briefly on the directive on undesirable substances and products of animal feed.
B:	i would now like to briefly to the directive on undesirable substances and products in the nutrition of them.

R:	it was then clearly shown that we can in fact tackle enlargement successfully within the eu ’s budget.
C:	at that time was clear that we can cope with enlargement, in fact, within the framework drawn by the eu budget.
B:	at that time was clear that we actually enlargement within the framework able to cope with the eu budget, the drawn.

Figure 1: Examples where annotator 1 judged the reordered system to give an improved translation when compared to the baseline system. Recall that annotator 1 judged 40 out of 100 translations to fall into this category. These examples were chosen at random from these 40 examples, and are presented in random order. **R** is the human (reference) translation; **C** is the translation from the system with reordering; **B** is the output from the baseline system.

R:	on the other hand non-british hauliers pay nothing when travelling in britain.
C:	on the other hand, foreign kraftverkehrsunternehmen fi gures anything if their lorries travelling through the united kingdom.
B:	on the other hand, fi gures foreign kraftverkehrsunternehmen nothing if their lorries travel by the united kingdom.
R:	i think some of the observations made by the consumer organisations are included in the commission ’s proposal.
C:	i think some of these considerations, the social organisations will be addressed in the commission proposal.
B:	i think some of these considerations, the social organisations will be taken up in the commission ’s proposal.
R:	during the nineties the commission produced several recommendations on the issue but no practical solutions were found.
C:	in the nineties, there were a number of recommendations to the commission on this subject to achieve without, however, concrete results.
B:	in the 1990s, there were a number of recommendations to the commission on this subject without, however, to achieve concrete results.
R:	now, in a panic, you resign yourselves to action.
C:	in the current paniksituation they must react necessity.
B:	in the current paniksituation they must of necessity react.
R:	the human aspect of the whole issue is extremely important.
C:	the whole problem is also a not inconsiderable human side.
B:	the whole problem also has a not inconsiderable human side.

R:	in this area we can indeed talk of a european public prosecutor.
C:	and we are talking here, in fact, a european public prosecutor.
B:	and here we can, in fact speak of a european public prosecutor.
R:	we have to make decisions in nice to avoid endangering enlargement, which is our main priority.
C:	we must take decisions in nice, enlargement to jeopardise our main priority.
B:	we must take decisions in nice, about enlargement be our priority, not to jeopardise.
R:	we will therefore vote for the amendments facilitating its use.
C:	in this sense, we will vote in favour of the amendments which, in order to increase the use of.
B:	in this sense we vote in favour of the amendments which seek to increase the use of.
R:	the fvo mission report mentioned refers specifi cally to transporters whose journeys originated in ireland.
C:	the quoted report of the food and veterinary offi ce is here in particular to hauliers, whose rushed into shipments of ireland.
B:	the quoted report of the food and veterinary offi ce relates in particular, to hauliers, the transport of rushed from ireland.

Figure 2: Examples where annotator 1 judged the reordered system to give a worse translation than the baseline system. Recall that annotator 1 judged 20 out of 100 translations to fall into this category. These examples were chosen at random from these 20 examples, and are presented in random order. **R** is the human (reference) translation; **C** is the translation from the system with reordering; **B** is the output from the baseline system.