

Improved HMM Alignment Models for Languages with Scarce Resources

Adam Lopez and Philip Resnik
University of Maryland

Overview

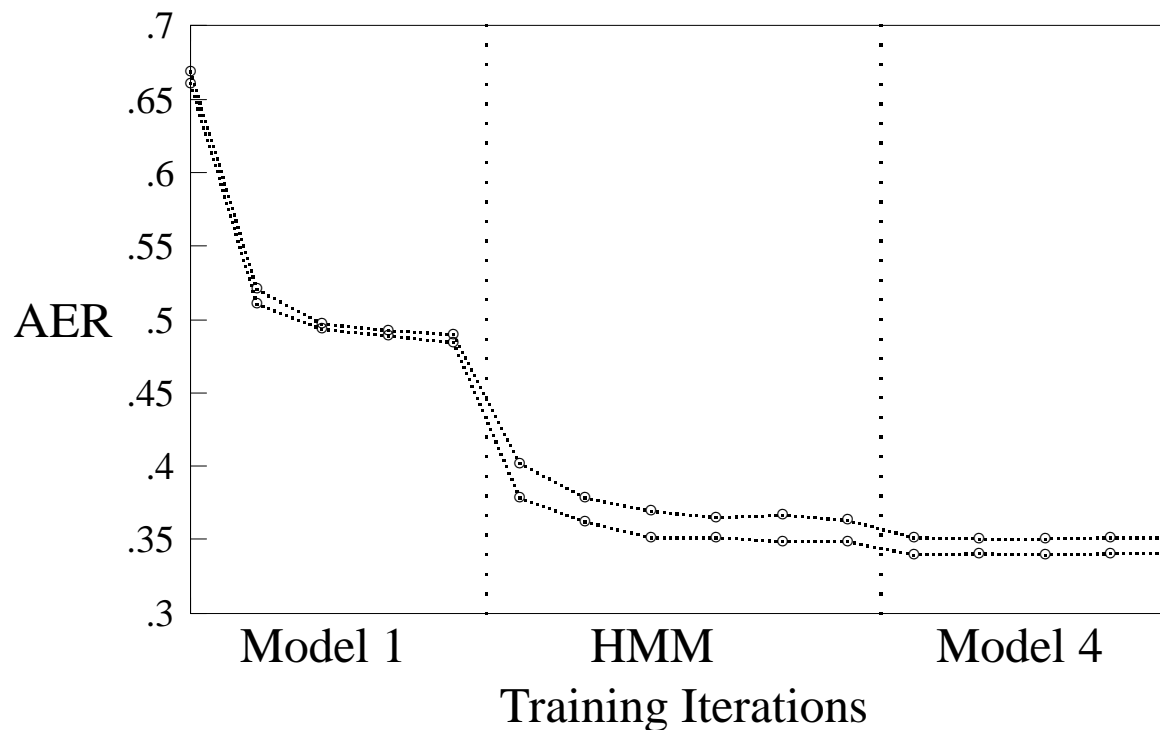
- Our approach is simple, flexible, and efficient
- Our framework can incorporate as much or as little annotation as we have available in either language
- We get good results on shared task with zero processing in scarce language (but we could use it if available)
- Talk outline:
 - Motivation
 - Model Outline
 - Results
 - Conclusions and Future Work

IBM Model 4: Benefits and Drawbacks

- Widely used (Yarowsky et al. 2001, *HLT*; Smith & Smith 2004, *ACL*; Och 2004, *CL*; Chiang 2005 *ACL*; Hwa et al. 2005, *NLE*)
- Good performance (*WPT* 2003; Och 2003, *CL*)
- Freely available implementations (Al-Onaizan et al. 1999, *JHU CLSP workshop*; Och 2003, *CL*)
- Problem: many local maxima, need good initial estimate for EM training
 - Solution: initialize using Model 1 and **HMM**
- Problem: search space cannot be efficiently enumerated
 - Solution: generate good initial alignment using Model 2 or **HMM**, partially expand with hill-climbing

Model 4 Alignment Performance

Learning curves for Romanian-English and English-Romanian, training sequence $1^5H^54^5$ (c.f. Och 2003, *CL*)



Observation: HMM is doing most of the work!

Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What would those things be ?

Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

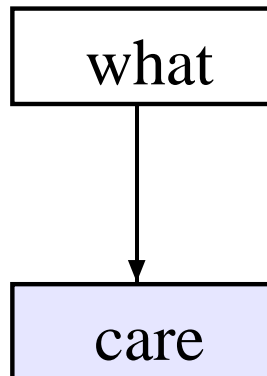
What would those things be ?

what

Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What would those things be ?

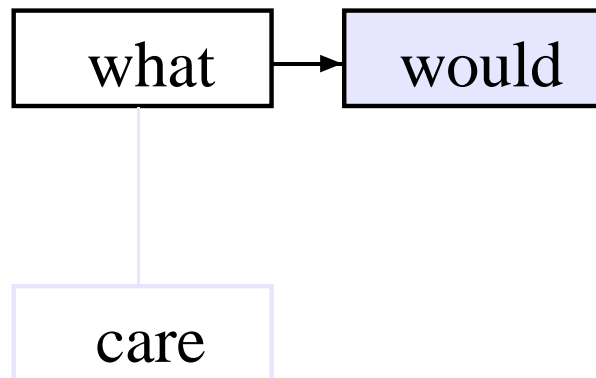


Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What **would** those things be ?

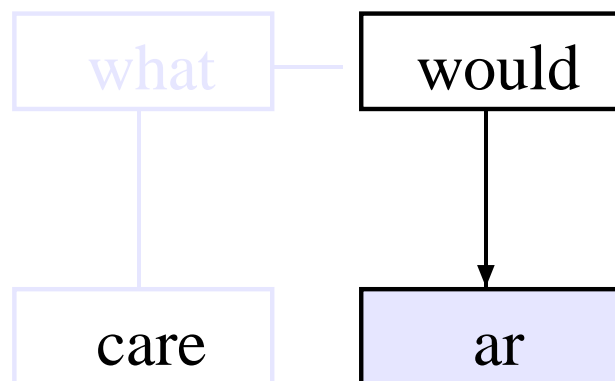
$$P(\textit{What} \longrightarrow \textit{would})$$



Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What **would** those things be ?

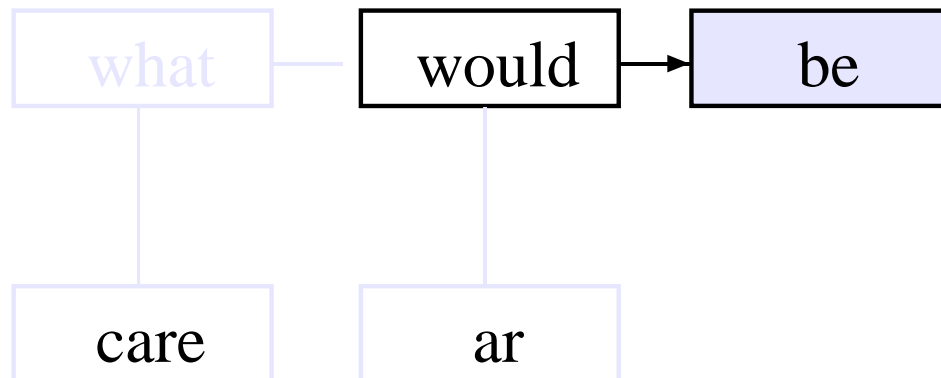


Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What would those things **be** ?

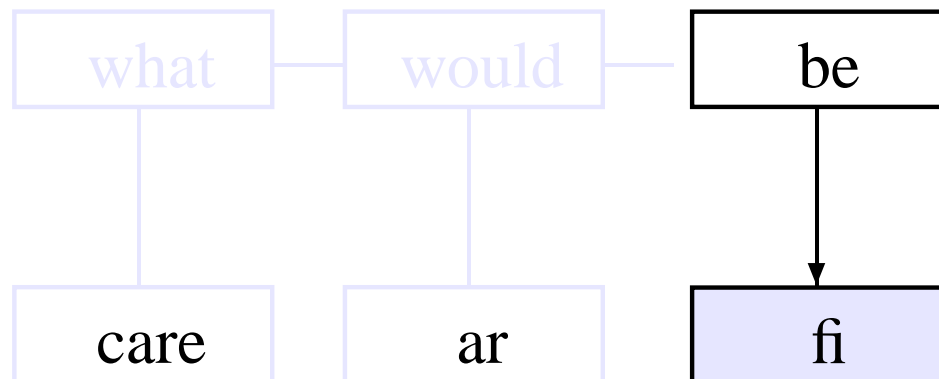
$$P(\textit{would} \longrightarrow \textit{be})$$



Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What would those things **be** ?

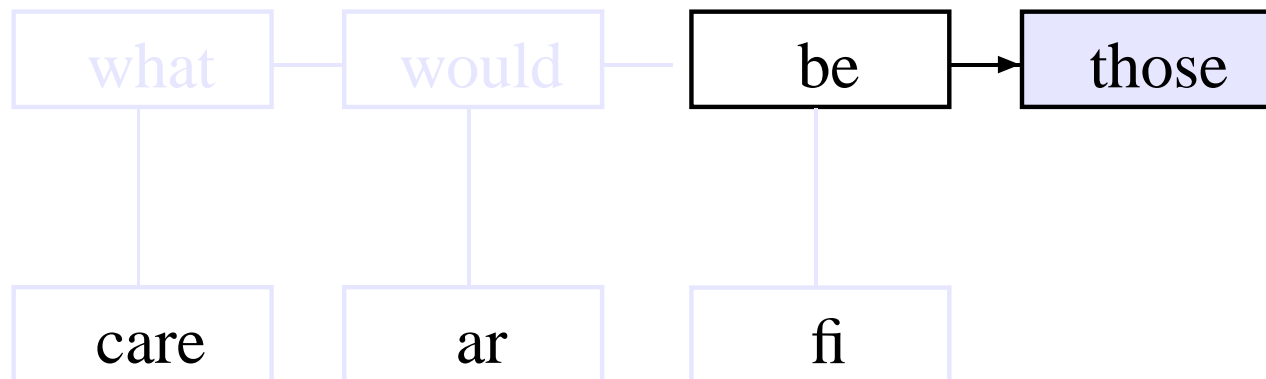


Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What would **those** things be ?

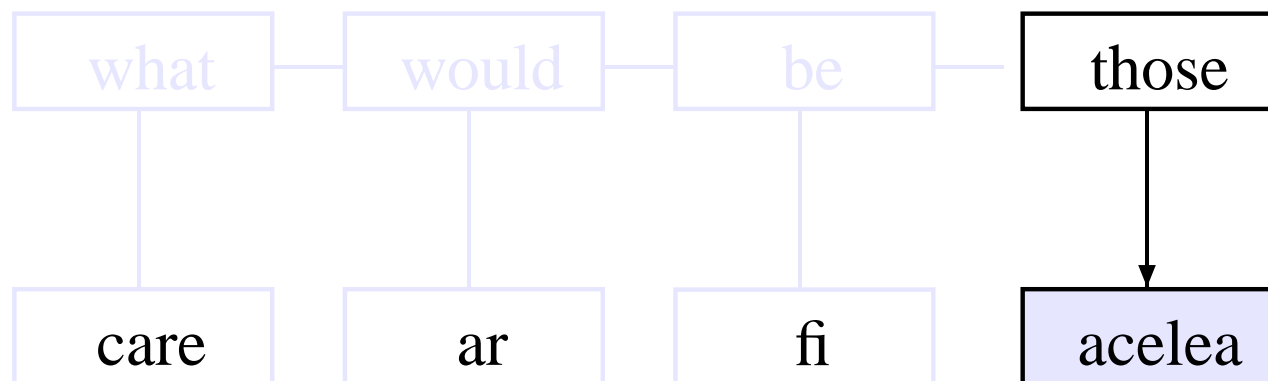
$$P(\text{be} \longrightarrow \text{those})$$



Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What would **those** things be ?

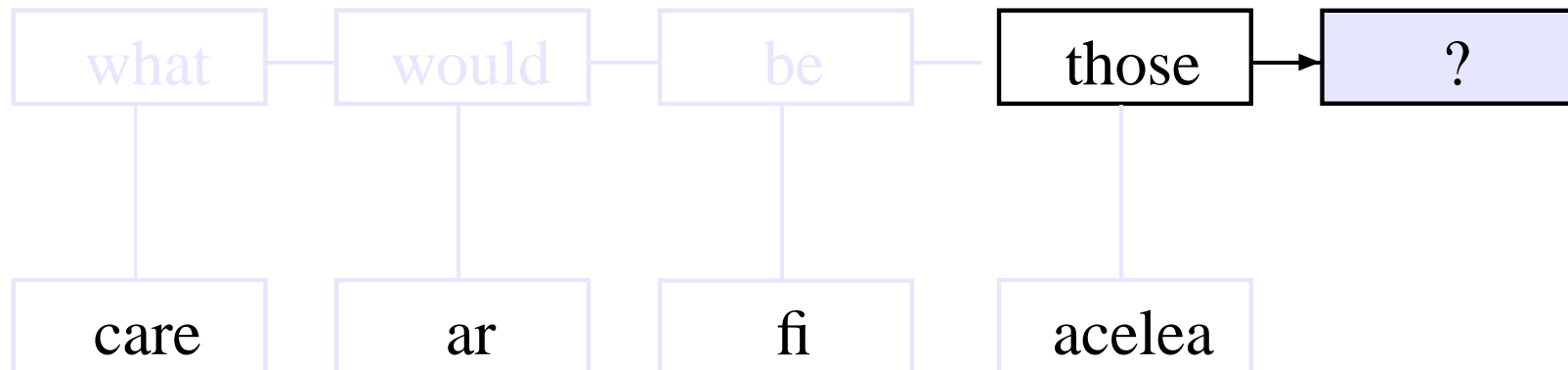


Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What would those things be ?

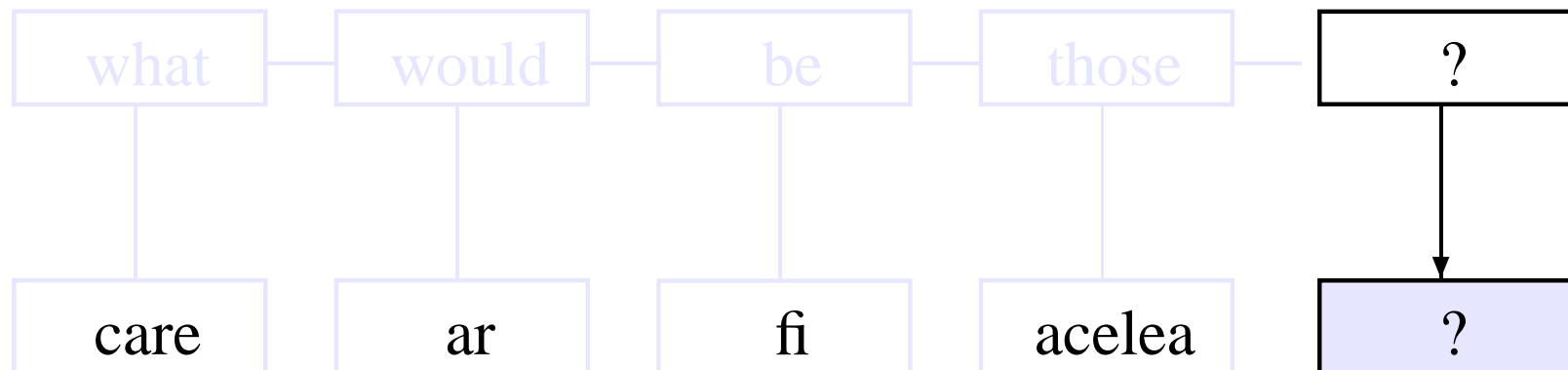
$P(\textit{those} \longrightarrow ?)$



Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

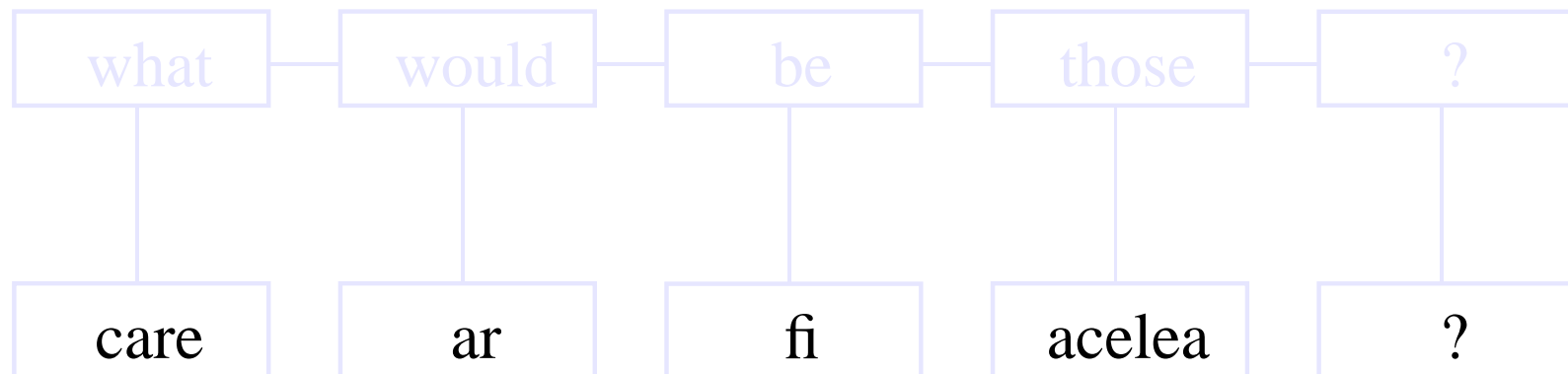
What would those things be ?



Idea: Improve HMM Model

- HMM formulation is much simpler than Model 4
 - Each English word is a state
 - Each Romanian word is an output symbol

What would those things be ?



Surface Distortion Model

Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class
(Och 1999, *EACL*)

What would those things be ?

Surface Distortion Model

Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)

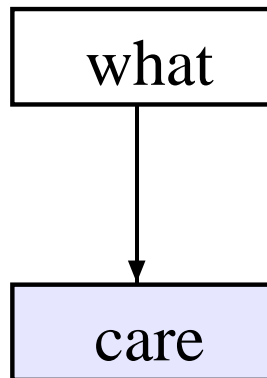
What would those things be ?

what

Surface Distortion Model

Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)

What would those things be ?

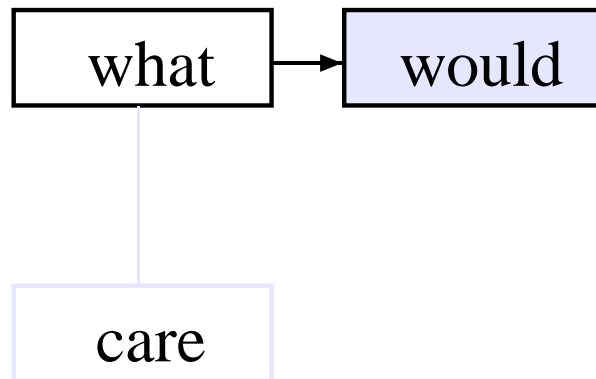


Surface Distortion Model

Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)

1
What **would** those things be ?

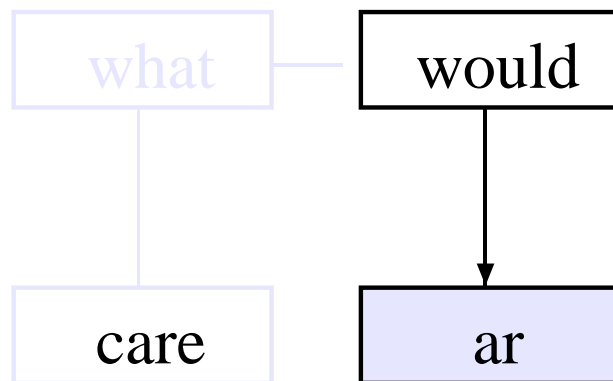
$$p(1|C(\textit{What}))$$



Surface Distortion Model

Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)

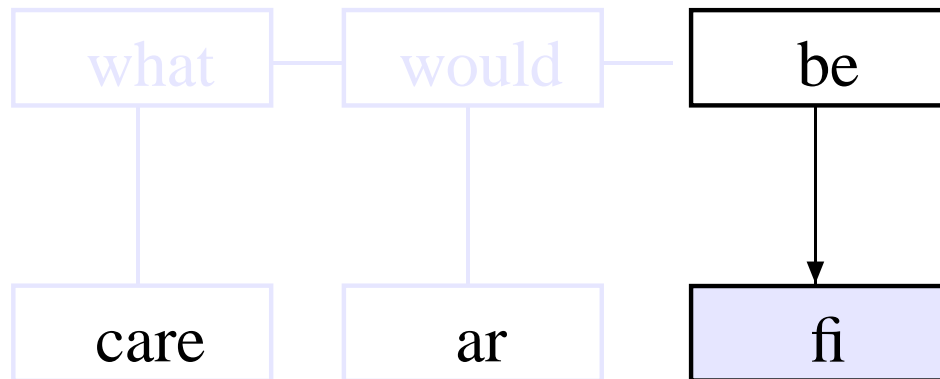
1
What **would** those things be ?



Surface Distortion Model

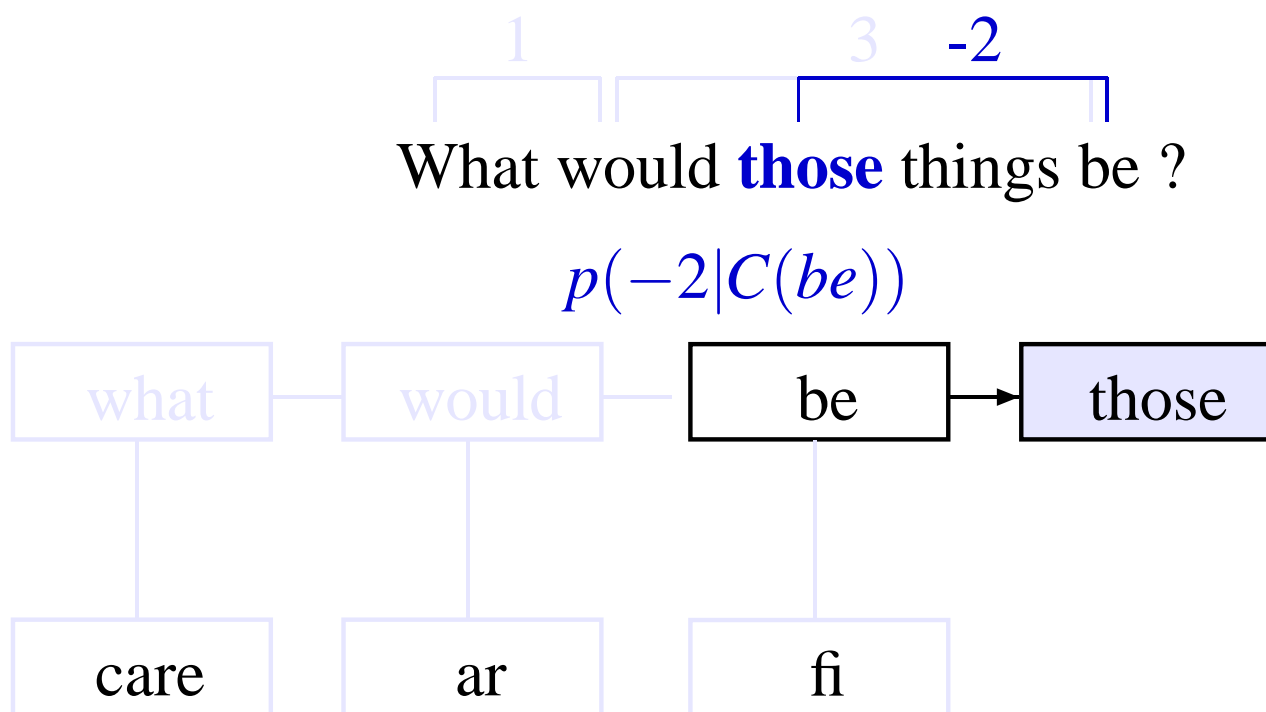
Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class
(Och 1999, *EACL*)

1 3
┌───┬──────────┐
What would those things **be** ?



Surface Distortion Model

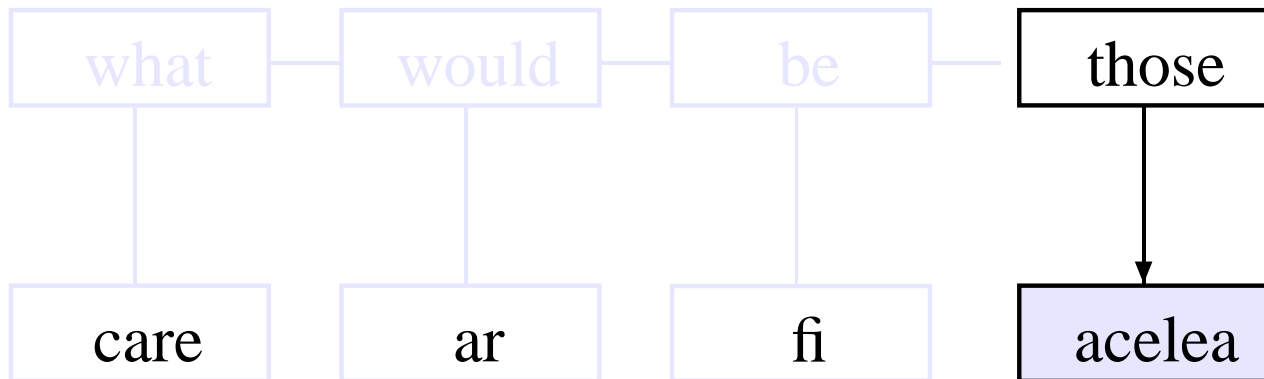
Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)



Surface Distortion Model

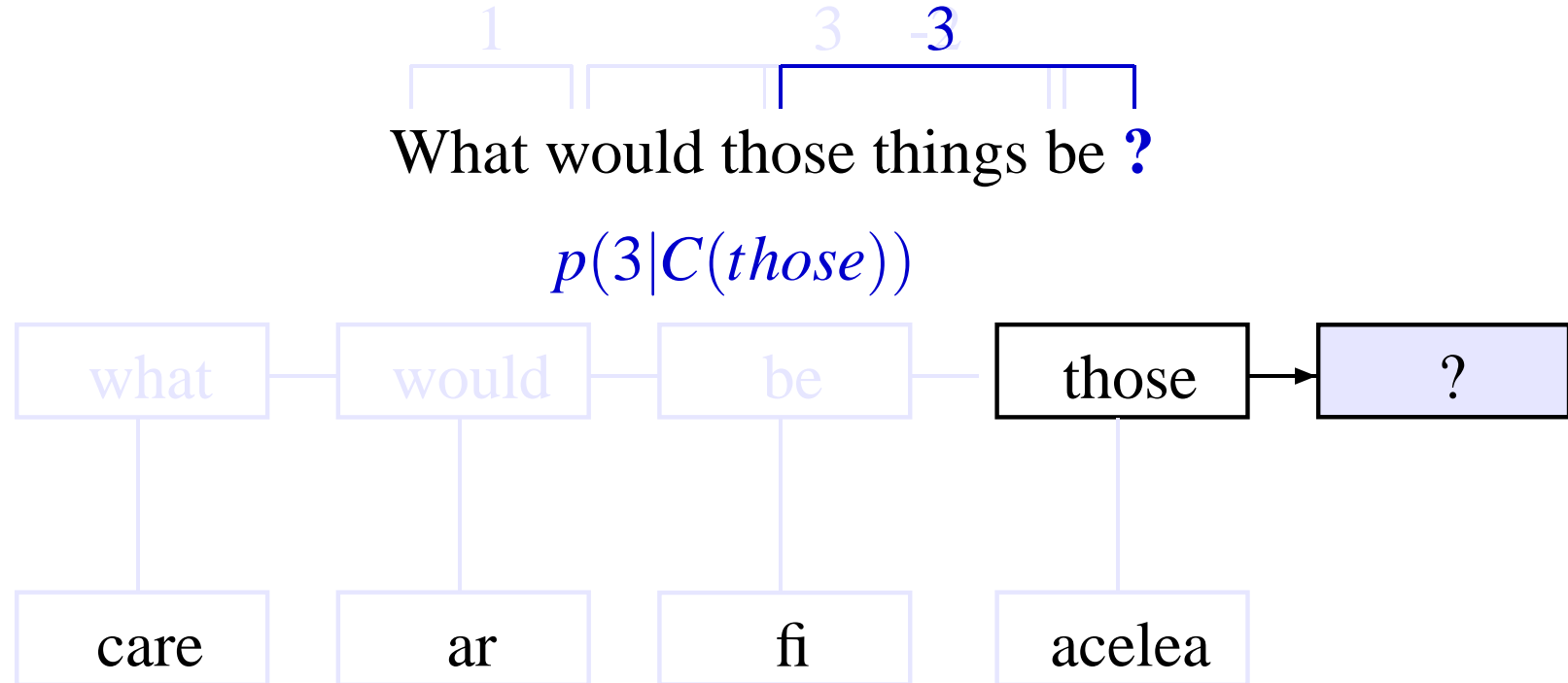
Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)

1 3 -2
What would **those** things be ?



Surface Distortion Model

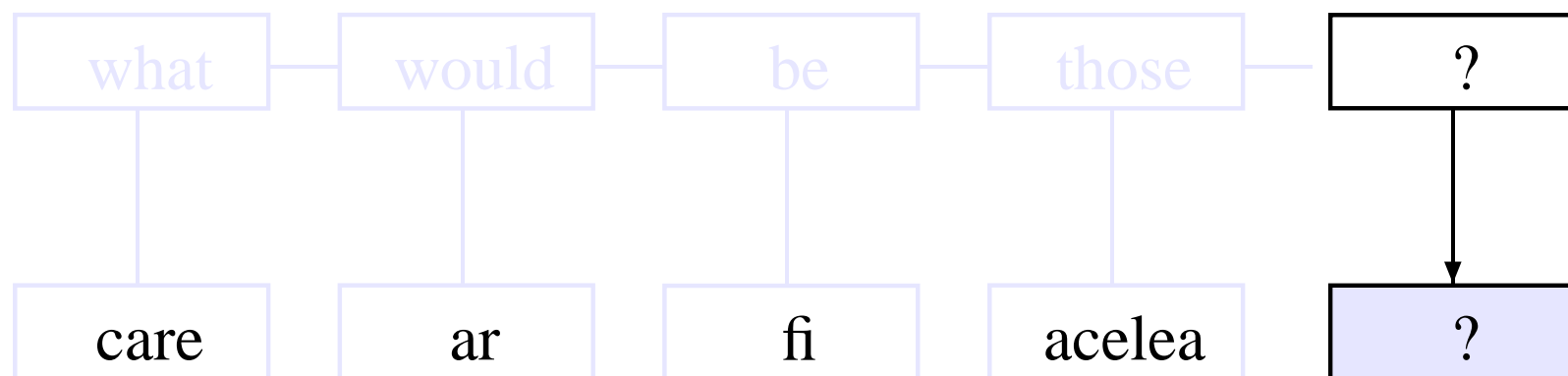
Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)



Surface Distortion Model

Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)

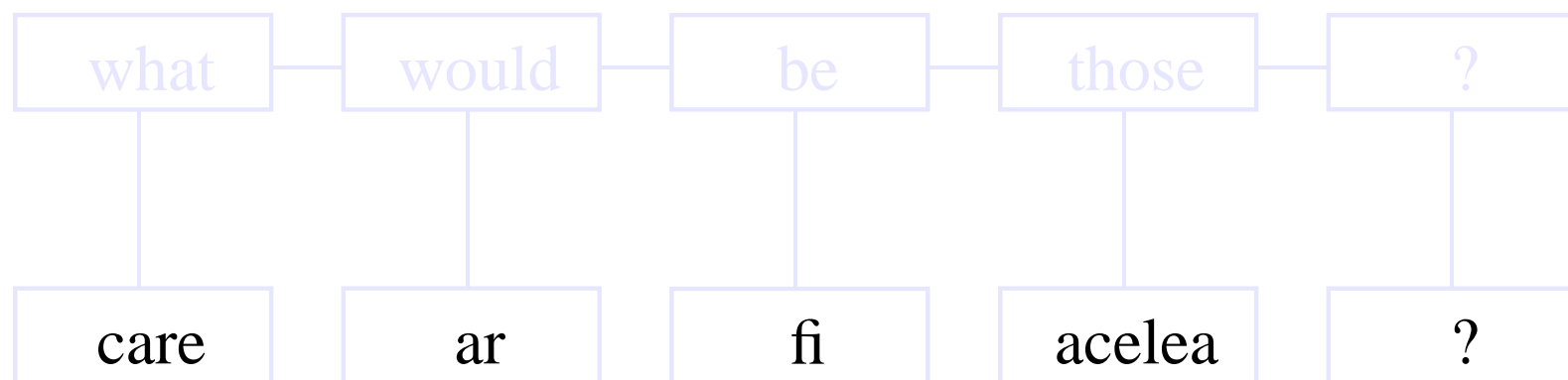
1 3 -3
What would those things be ?



Surface Distortion Model

Parameterize transition probability on the surface distance between words, conditioned on an unsupervised word class (Och 1999, *EACL*)

1 3 ~~3~~
What would those things be ?




Question: Can We Encode Syntax in Our HMM?

- Much interest in syntax-based alignment and translation models
- We *don't* expect to have syntax for scarce languages
- We *do* have syntax for English
- We *don't* want to be tied to any particular syntactic model
- So: we need a framework that uses whatever syntactic analysis we have, but doesn't depend on it

Idea: Tree Distortion Model


- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction



What would those things be ?

Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction




What would those things be ?

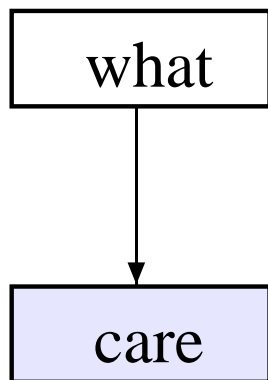
what

Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction

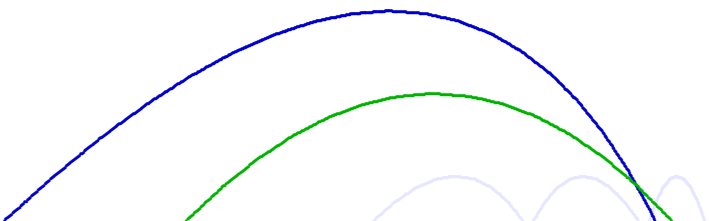


What would those things be ?



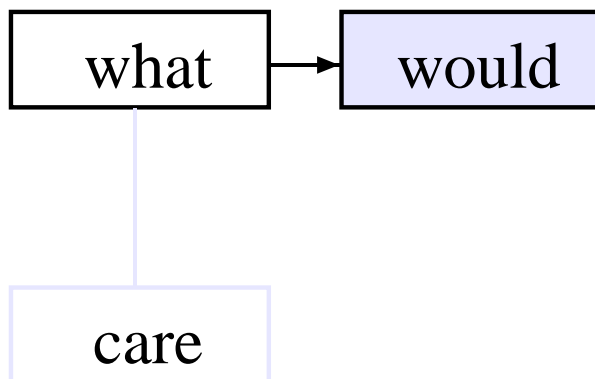
Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction



What **would** those things be ?

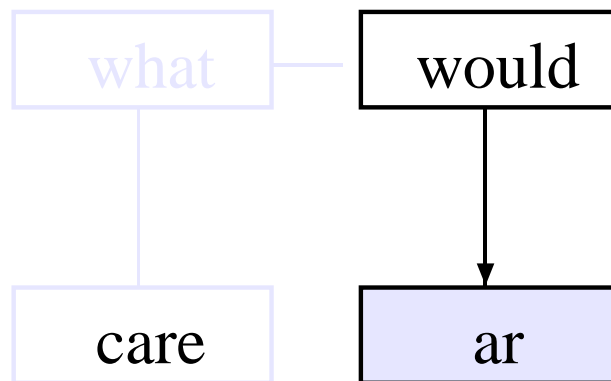

$p(1 \text{ up}, 1 \text{ down}, \text{to the left} \mid T(\textit{What}))$



Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction

What **would** those things be ?

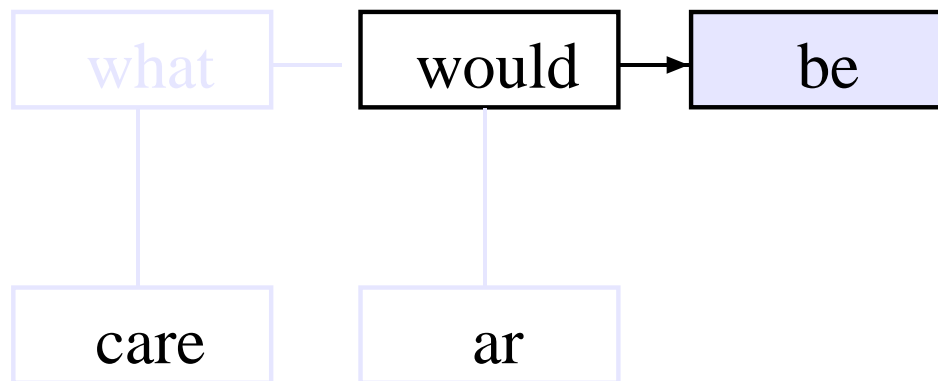


Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction

What would those things be ?


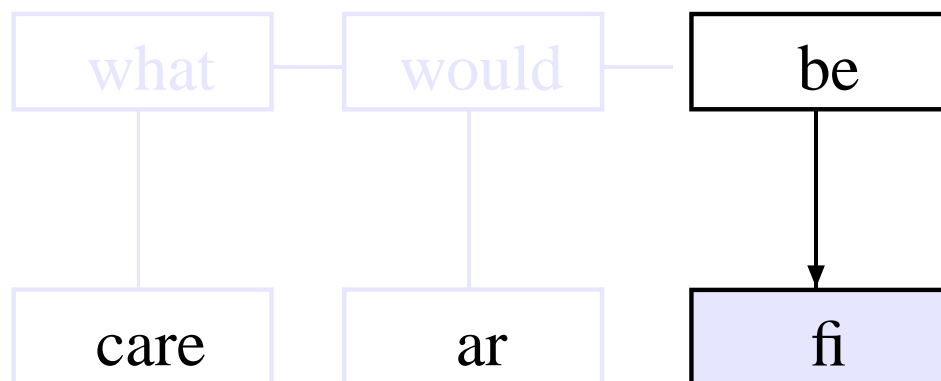
$p(1 \text{ up}, 0 \text{ down, to the left} | T(\textit{would}))$



Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction

What would those things **be** ?

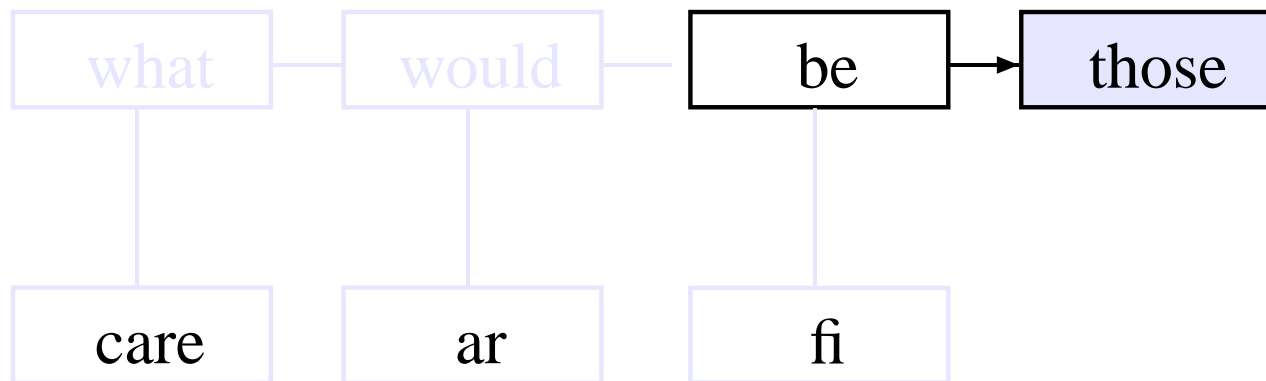
The diagram shows the sentence "What would those things be ?" with dependency arcs. A large arc connects 'be' to 'what'. A medium arc connects 'be' to 'would'. A small arc connects 'be' to 'care'. A very small arc connects 'be' to 'ar'. A tiny arc connects 'be' to 'fi'.

Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction

What would **those** things be ?


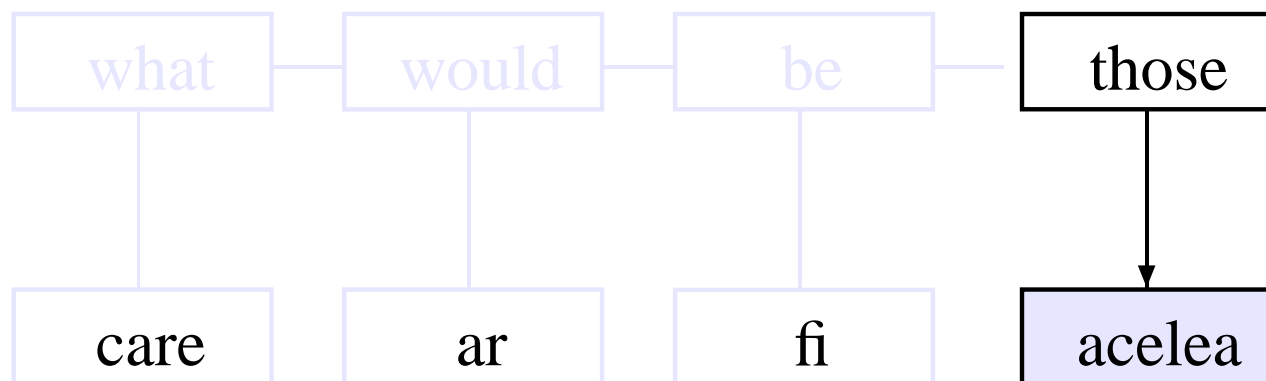
$p(0 \text{ up, } 2 \text{ down, to the right} \mid T(\text{be}))$



Idea: Tree Distortion Model

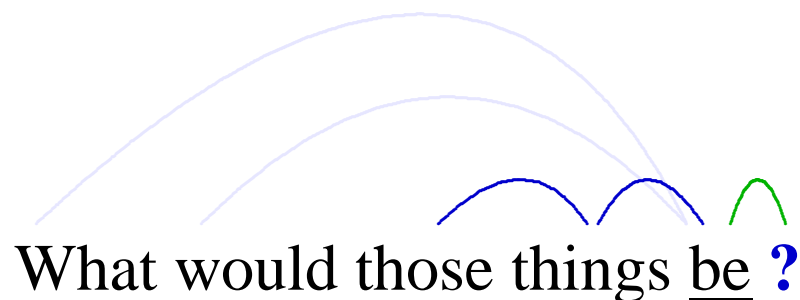
- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction

What would **those** things be ?

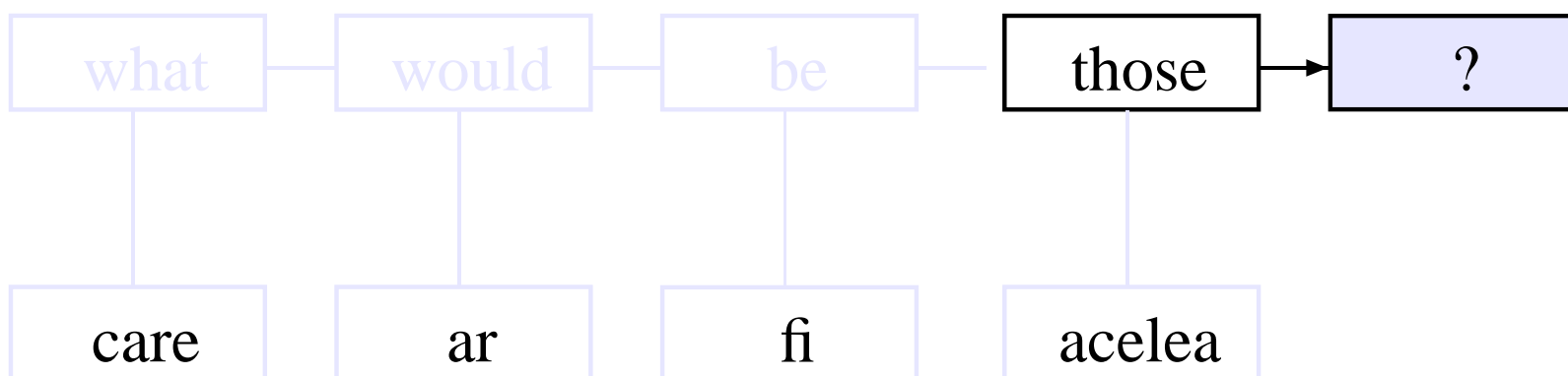
A diagram illustrating dependency arcs. The word 'those' is at the top right. Four arcs originate from it: one to 'what' (top left), one to 'would' (top middle), one to 'be' (top right), and one to 'acelea' (bottom right). The arcs are drawn as curved lines above the words.

Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction




$p(2 \text{ up}, 1 \text{ down}, \text{to the left} \mid T(\text{those}))$

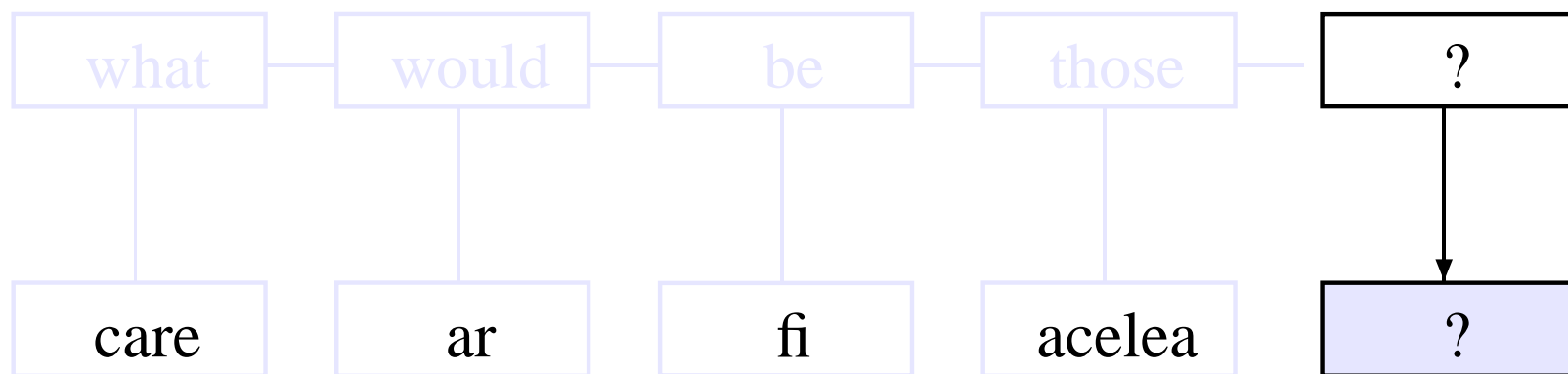


Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction




What would those things be ?

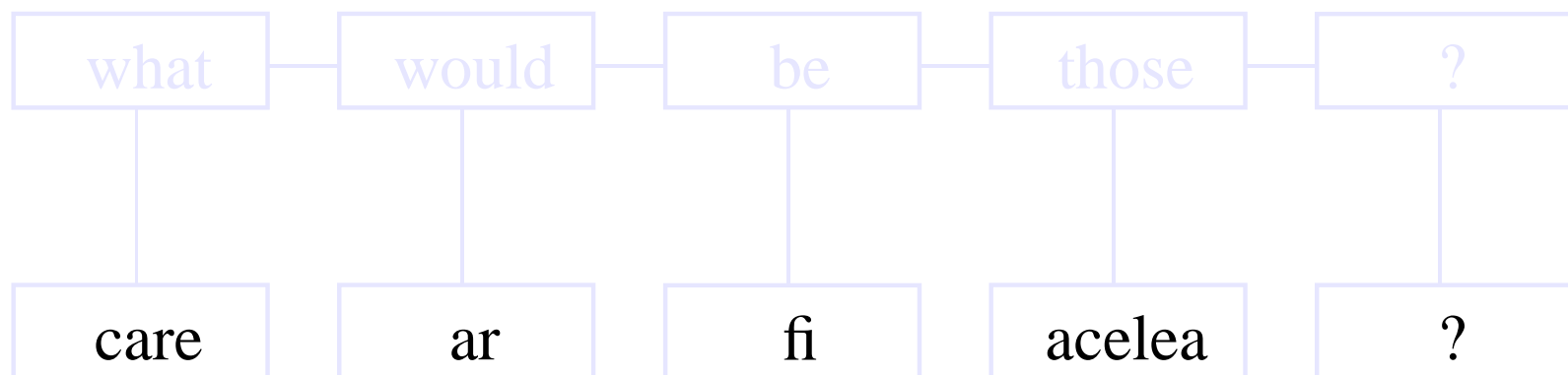


Idea: Tree Distortion Model

- Parameterize transition probability on the *tree distance* between words, conditioned on part-of-speech
- Tree distance: number and orientation of dependency links traversed + direction



What would those things be ?



Tree Distortion Model (Continued)

- Corresponds naturally to concepts such as postmodifier (0 up, 1 down, left)
- Tree distortion can be converted to use with CFG, TAG parsing models – not tied to dependency models
- Can use tree distortion for English-Romanian; surface distortion for Romanian-English
- Even better: since surface distortion and tree distortion model the same thing, we can combine them using linear interpolation
- Interpolation coefficients trained with other model parameters using EM

Question: What's the Impact of Better Initialization?

- Usual approach:
 - Initialize translation probabilities from IBM Model 1, 5 iterations
 - Initialize distortion probabilities uniform
- Our approach:
 - Initialize translation probabilities using smoothed log-likelihood ratio similar to Moore (2004, *ACL*), 2 iterations
 - Initialize distortion probabilities with a bias towards short surface distances

Question: Does Null Alignment Help?

- Null alignment used in many alignment models, including HMM (Och, 2000, *COLING*)
- *However*, we are going to symmetrize our alignments
- Symmetrization usually based on intersection (Och, 2000, *ACL*, Koehn et al. 2003, *NAACL*)
- Null alignment and intersection both accomplish the same thing: improve precision, cost recall
- We don't permit null alignment. Expected result:
 - Alignment step maximizes recall
 - Symmetrization step maximizes precision

Results

HMM unlimited uses tree distortion; HMM limited does not

		Precision	Recall	AER
Inuktitut	Model 4 ($1^5H^54^5$)	.8682	.5700	.2801
	HMM lim.	.8916	.6280	.2251
Romanian	Model 4 ($1^5H^54^5$)	.7620	.5134	.3865
	HMM lim.	.7377	.6169	.3281
	HMM unlim.	.7241	.6215	.3311

Reduction in AER: Inuktitut (19.6%), Romanian (15.1%)

Discussion

- Improved recall from our null alignment hypothesis
- Improved precision from our initializer
- Our system gets best results among those that do zero processing in the scarce language
- Our framework could incorporate the features from best systems
 - morphology, use in classifier combination (Schafer & Drábek 2005, *WPT*)
 - morphology approximations and minimum error rate training (Fraser & Marcu 2005, *WPT*)
 - dictionaries, cognates, and other analyses (Aswani & Gaizauskas 2005, *WPT*; Tufis et al. 2005, *WPT*)

Conclusions

- We can produce state-of-the-art results with HMM
- We can easily combine complementary reordering models using linear interpolation; we could also do this for word-to-word translation models, e.g. using part-of-speech (Toutanova et al. 2002 *EMNLP*) or hierarchical models (Nießen & Ney 2004, *CL*)
- Initialization is *really* important!
- Null alignment is not so important. Asymmetry + intersection approach allows us to exploit precision / recall tradeoff
- If we want to use Model 4, we can still incorporate our HMM as an improved initializer

This slide intentionally left blank

Formulae

- HMM: $P(f_1^J | e_1^I) = \prod_{j=1}^J \sum_{i=1}^I \sum_{i'=1}^I P(a_j = i | a_{j-1} = i') \cdot P(f_j | e_i)$
- surface distortion: $P(a_j | a_{j-1}) = p(a_j - a_{j-1} | C(e_{a_j}))$
- tree distortion: $P(a_j | a_{j-1}) = p(\tau(a_j, a_{j-1}) | T(e_{a_j}))$
- combined distortion:
 $\lambda_{C,T} p(a_j - a_{j-1} | C(e_{a_j})) +$
 $(1 - \lambda_{C,T}) p(\tau(a_j, a_{j-1}) | T(e_{a_j}))$

French and English Results

		Precision	Recall	AER
French	Model 4	.9730	.8058	.1094
	HMM lim.	.9566	.8284	.1028
	HMM unlim.	.9605	.8304	.0999
Chinese	Model 4	.7200	.4810	.4192
	HMM lim.	.7191	.6311	.3256
	HMM unlim.	.7114	.6387	.3251

Reduction in AER: French (8.6%), Chinese (19.6%)