Decoder-Guided Backoff

Using Word Lattices to Improve Translation from Morphologically Complex Languages

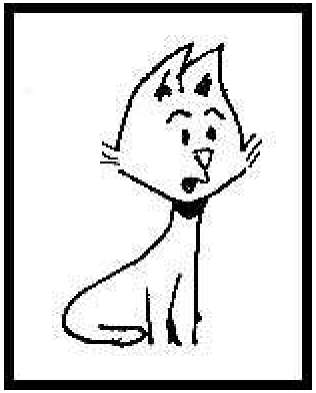
Chris Dyer
University of Maryland

Outline this talk

- What is morphology and why does it matter to MT?
- Prior work
- Modeling morphology as observational ambiguity
- Decoding word lattices
- Experimental results

What is morphology? A crash course in words

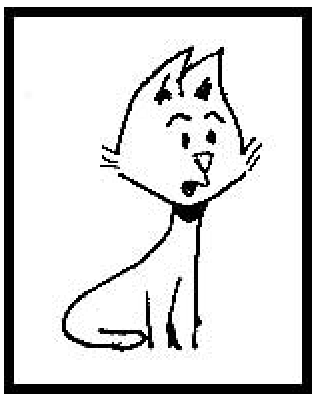
- An important observation: words have complex internal structure.

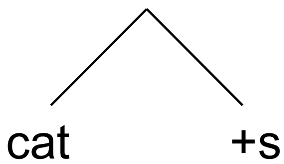


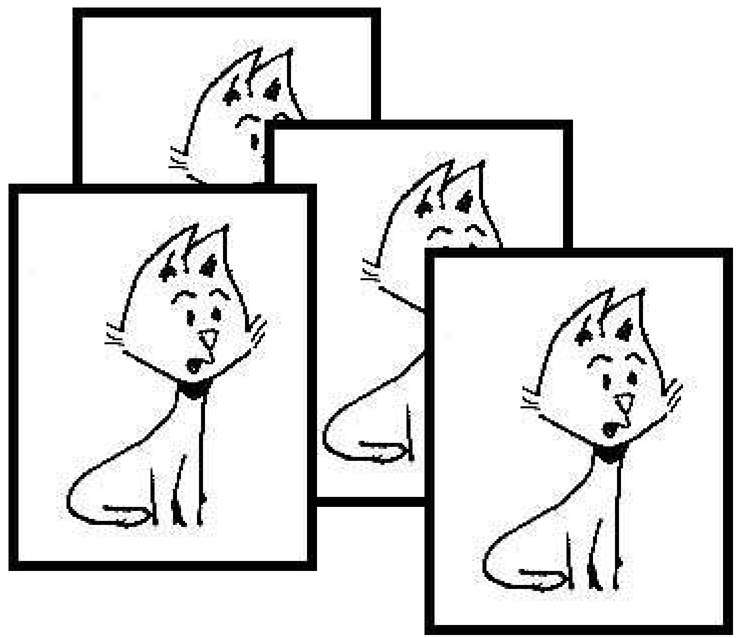
cat

What is morphology?
A crash course in words

- An important observation: words have complex internal structure.







Morphology

- Conventional division:
  - ***Derivational morphology***
    - "Derive" new forms from a root
    - Adjective → Verb (wide → widen)
    - Verb → Noun (destroy → destruction)
  - ***Inflectional morphology***
    - "Add meaning" to a base category
    - +PLURAL (cat → cats)
    - +DATIVE (der Student → dem Studenten)
    - +FUTURE (ser → será)

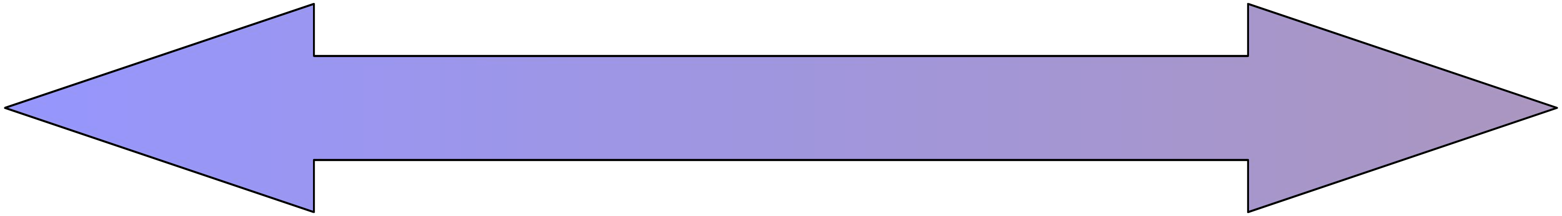
Morphology

- Clitics
  - Some words attach to other words.
  - But, orthographic conventions differ:
    - the boy
    - ***al***walad (the boy)

    - She hit him.
    - darabat***hu***. (She hit him.)

A field guide to morphology

Analytic/Isolating — Synthetic



Chinese	English	Spanish	Czech	Maltese	Turkish	Navaho
		Italian	Polish	Arabic	Finnish	Inuktitut
		French	Russian	Hebrew	Hungarian	Mohawk
			Welsh		Basque	
			Irish			
			German			
			Danish			

Analytic languages

- No inflectional (category-preserving) morphology
- Some derivational (esp. compounding) morphology

明天	我	的	朋友	为	我	做	生日	蛋糕
míngtīan tomorrow	wǒ I	de 's	péngyou friend(s)	wéi for	wǒ I	zuò to make	shēngrì birthday	dàngāo cake

"My friends will make me a birthday cake tomorrow."

Fusional languages

- Fusional
  - Most Indo-European languages.
  - Many functional morphological elements (eg. tense, number, gender) combined into a single morpheme.
    - She sing**s**. +s = singular, present tense, indicative

Agglutinative languages

- Agglutinative
  - Hungarian, Finnish, Turkish
  - Concatenate chains of (mostly *functional*) morphemes

Uygar-laş-tır-a-ma-dık-lar-ımız-dan-mı-sınız?

Civilized-VERB-CAUS-ABLE-NEG-NOM-PLU-POS1P-ABL-INT-2PL.AGR

“Are you from the ones we could not civilize?”

Polysynthetic languages

- One word, many morphemes

  aliiku-sersu-i-llammas-sua-a-nerar-ta-ssa-galuar-paal-li

  "*However, they will say that he is a great entertainer.*"

- A single word may include several open- and closed- class morphemes

aliiku = entertainment	*a = say*
sersu = provide	*llamas = good at*

Morphology & MT

- So why, as MT researchers, do we care about morphology?

  1. Inflectional richness → free word order

  2. Data sparseness

Morphology & MT

- So why, as MT researchers, do we care about morphology?

  1. Inflectional richness → free word order

  2. Data sparseness

Prior work

- Goldwater & McClosky (2005)
  - Czech → English
  - Preprocess the corpus to throw away some morphemes:
    - Word truncation (ask F.J. Och)
    - Lemmatize everything
    - Only lemmatize infrequent words
    - Keep inflectional morphemes that “mean something” in English
  - Experimentation necessary to determine best process!

Prior work

- Goldwater & McClosky (2005) results:

	Dev	Test
word-to-word	.311	.270
lemmatize all	.355	.299
except Pro	.350	
except Pro, V, N	.346	
lemmatize $n < 50$	.370	.306
truncate all	.353	.283

*BLEU scores with 5 reference translations, *word*-based SMT system.

Prior work

- However, with a phrase-based translation model and more data, things look a bit different:

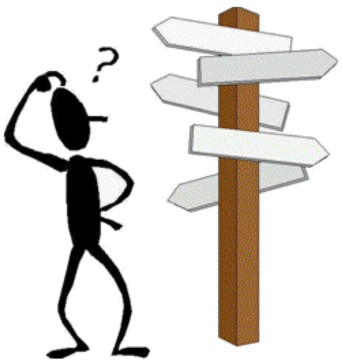
Input	BLEU*
Surface	22.81
Truncated (l=6)	22.07
Lemmas	22.14

p<.05

*** 1 reference translation, WMT07 dev-test**

Prior work

- What happened?
  - The morphemes that were thrown away had useful information
  - Must avoid *two* pitfalls



Data Sparseness

Information Loss

A Better Translation

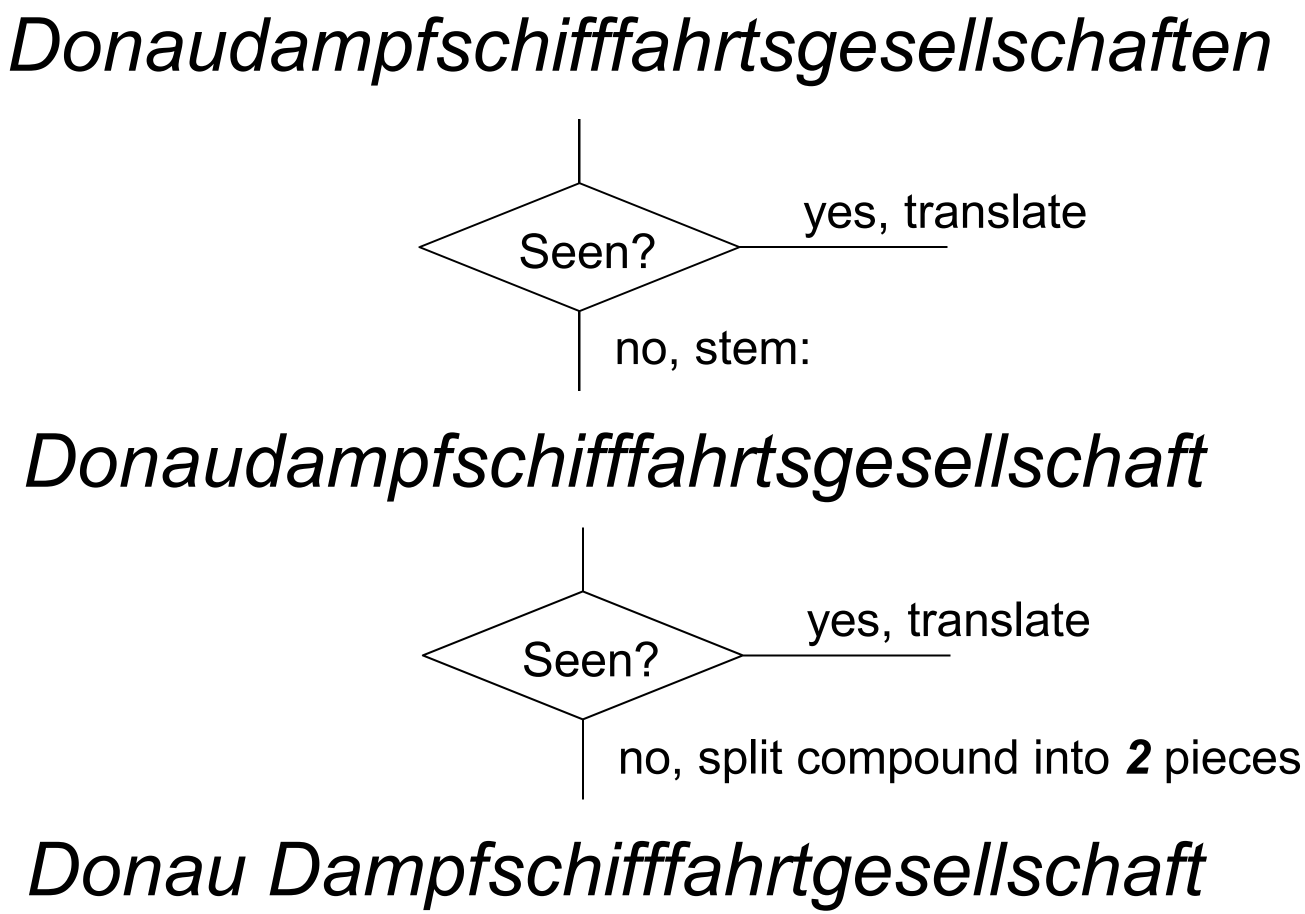
Prior work

- Talbot and Osborne (2006)
  - Learn “redundancies” automatically from a parallel corpus
  - Only collapse distinctions that are meaningless w.r.t. a particular target language
  - Experiments
    - Smooth surface translation table with revised probabilities
    - Use “compressed” lexicon just to improve word alignments

Prior work

- Yang & Kirchhoff (2006)
  - Backoff models for machine translation
  - If you don't know how to translate a word, perform morphological simplification
  - Experiments on Finnish & German
    - German
      - fusional morphology
      - productive compounding
    - Finnish
      - agglutinative morphology
      - Limited noun-noun compounding

Prior work: Yang & Kirchhoff (2006)



Yang & Kirchhoff (2006)

GERMAN		
Training data	**baseline**	**backoff**
5k	15.3	16.3
50k	20.3	20.7
751k	24.8	25.1
FINNISH		
Training data	**baseline**	**backoff**
5k	12.9	14.0
50k	15.6	16.4
751k	22.0	22.3

Prior work: Yang & Kirchhoff (2006)

- Potential Problems
  - Everything is done as preprocessing
  - Only back off if $C(f) = 0$
  - No improved word alignment

Prior work: take-away

- Morphological simplification can help.
- Morphological simplification can hurt.
  - Only collapse meaningless distinctions!
  - Use a backoff strategy!
- All approaches presented involve making decisions about the translation forms in advance of decoding.
  - Question: **Is this the best strategy?**

Spoken Language Translation

- Recognize speech in the source language
  - ASR is not perfect!
- Translate into English
  - Translation is not perfect!

- Can we minimize error compounding?

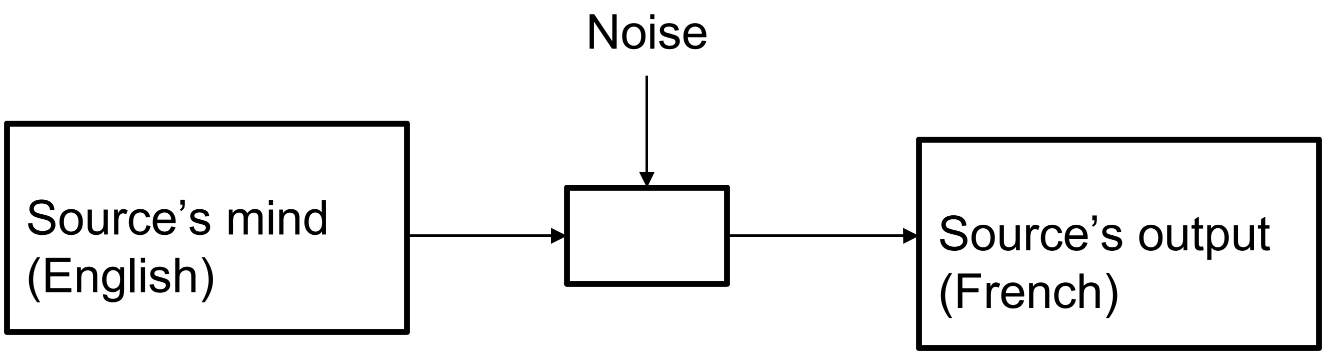
What SLT research tells us

- Joint models better perform better than translating the 1-best hypothesis
  - Ney (1999), Bertoldi et al. (2005a, 2007), Shen et al. (2006)
- Enumerating all hypotheses is not necessary
  - Confusion networks in phrase-based decoders (Moses), Bertoldi (2005a), Bertoldi et al. (2007)
  - Confusion networks in hierarchical (SCFG) decoders, Dyer & Resnik (2007)

Idea

Model the backoff problem to make it look like speech translation.

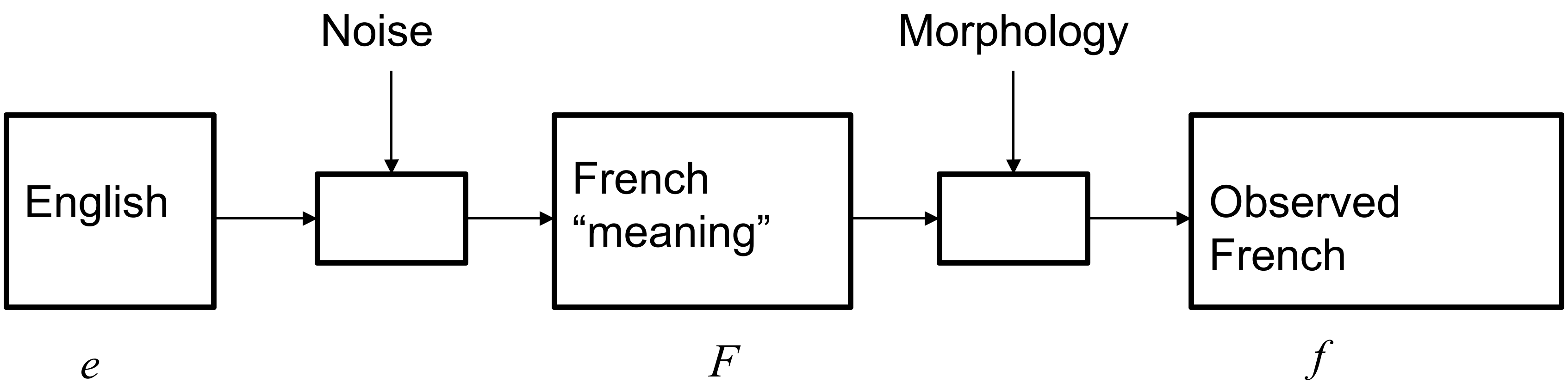
The noisy channel



Decoding:

$$\arg\max_e P(e \mid f) = \arg\max_e P(f \mid e) P(e)$$

A noisier channel



Approximation:

$$S(f) \approx F$$

Decoding:

$$\arg\max_{e} \max_{f' \in S(f)} P(e, f' \mid f)$$

Constructing a translation system

- What is $S(f)$?
  - Set of sentences
    - All morphological “alternatives” to f that the system might know how to translate
  - Cost function from a sentence to some value
    - ~How much information did we throw away?
- Constructing $S(f)$
  - Use existing morphological analyzers
  - Truncation
  - Compound splitting

Example

- Given the observed Spanish sentence: *la mujer vieja*, $S(f)$ might contain:

SENTENCE	PENALTY
la mujer vieja	*?*
EL mujer vieja	*?*
la mujer VIEJ	*?*
EL mujer VIEJ	*?*

Example

- What to do with the penalty?
  - Posterior probability of the sentence under some model (e.g. ASR/OCR word lattices)
  - Amount of morphological information thrown away
    - Count
    - Quantified under some model (e.g. Talbot & Osborne 2006)
  - Function of $\#(f)$ vs. $\#(g(f))$ in the training corpus

Representing *S(f)*

- *S(f)* is a huge list with scores! We'd like a compact representation of a huge list.
- Start simple: inflectional morphology
  - Single stem affected
- Confusion networks
  - Good at representing alternatives at a given position
  - Plus, we know how to decode them!

Czech-English translation

- Czech is a highly inflected fusional language.
- Not much compounding.

Language	*Tokens*	*Types*	*Singletons*
Czech	**1.2M**	**88037**	**42341**
cz-lemmas*	"	**34227**	**13129**
cz-truncated	"	**37263**	**13039**
English	1.4M	31221	10508
Spansh	1.4M	47852	20740
French	1.2M	38241	15264
German	1.4M	75885	39222

*** J. Hajič and B. Hladká. 1998. Tagging Inflective Languages.**

Confusion networks

- CN representation of *S(f)*
  - Surface and lemma at each position
  - Simple penalty model: surface=0, lemma=1

z	amerického	břehu	atlantiku	se	veskerá	taková	odůvodnění	jeví	jako	naprosto	bizarní	.
	americký	břeh	atlantik									

atlantiku

atlantik

Estimating a translation model

- $S(f)$ contains sentences that are a mixture of lemmas and surface forms
- Need translation model that contains both

Estimating a translation model

- Simple solution:
  - Train independent models in parallel
    - Surface → Surface
    - Lemma → Surface
  - Then merge or have two phrase tables available
  - Decoder to chooses the path/translation it likes best
  - Pros: easy to estimate
  - Cons: except within limits,mixed phrases do not exist!
- A variety of other model possibilities exist!

Czech-English results

Input	BLEU*
Surface forms only	22.74
Backoff (~Y&K '06)	**23.94**
Lemmas only	22.50
Surface+Lemma (CN)	**25.01**

- Improvements are significant at p<.05; CN > surface at p<.01.
- WMT07 training data (2.6M words), trigram LM

*** 1 reference translation**

Czech-English results

Surface only:

From the US side of the Atlantic all such odůvodnění appears to be a totally bizarre.

Lemma only:

From the [US] side of the Atlantic with any such justification seem completely bizarre.

Confusion Net (Surface+Lemma):

From the US side of the Atlantic all such justification appears to be a totally bizarre.

Representing other forms of ambiguitiy

- CNs are fine for inflection, but what about a language with compound/clitic splitting?

gesamthaushaltsplans

gesamthaushaltsplan

gesamt haus halt plans

gesamt haus halt plan

Different lengths!

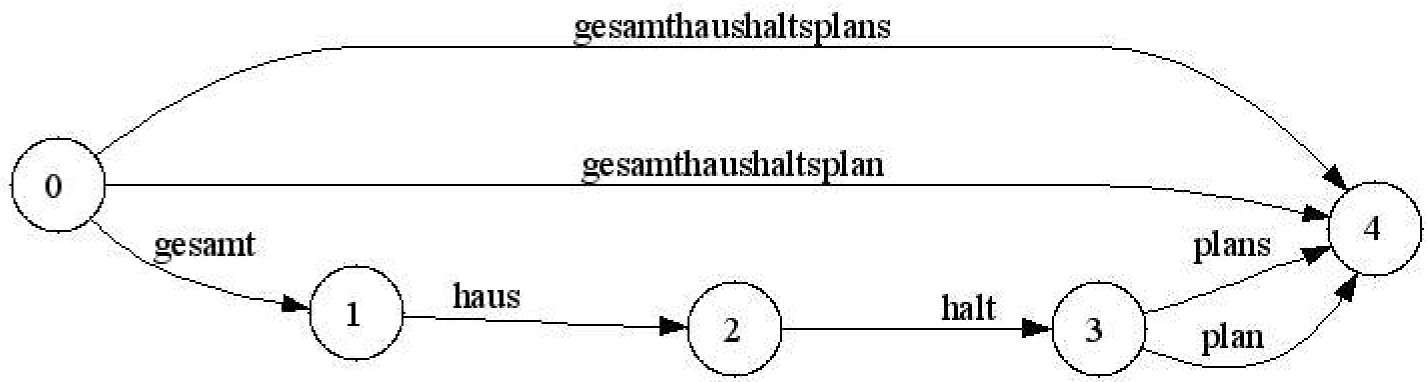
Confusion nets: the problem

- Every path must pass through every node

gesamthaushaltsplans	ε	ε	ε
gesamthaushaltsplan	haus	halt	plans
gesamt			plan

Word lattices

- Any set of strings can be represented
- Algorithms exist for minimizing their size



Decoding word lattices I: Create a chart from the lattice*

- Number nodes by distance from start-node
- For each edge leaving node i and labeled with word w, place word w into column i
- Augment cell with *span length* (difference between number of next node and current node)

gesamthaushaltsplans **4**	haus **1**	halt **1**	plans **1**
gesamthaushaltsplan **4**			plan **1**
gesamt **1**			

* Based on a CKY parser for lattices by Cheppalier (1999)

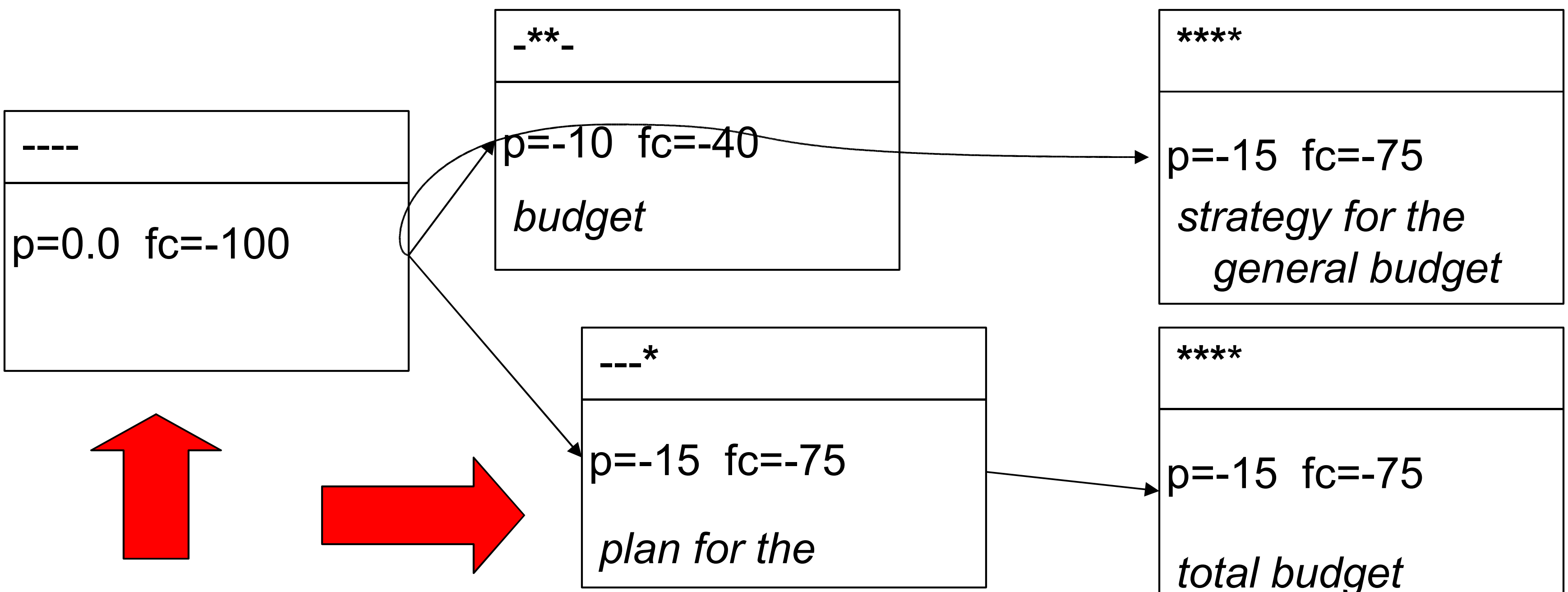
Decoding word lattices II

- Create translations options for column spans (rather than word spans)
- Column coverage replaces word coverage
- Search for a hypothesis that covers all columns.

  A word may span more than one column!

Decoding word lattices III

gesamthaushaltsplans **4**	haus **1**	halt **1**	plans **1**
gesamthaushaltsplan **4**			plan **1**
gesamt **1**			



Word lattice decoding: Problems

- The standard exponential decay distortion model is very poorly defined for word lattices!
  - Lexicalized reordering models fare better.
- Span limits are also poorly defined.

Efficiency of word lattice decoding

- "Morphology" lattices are compact
  - Many nodes that all paths pass through (quasi-linear networks)
  - ASR word lattices do not necessarily have this property!
- Running time proportional to the length of the longest path

Efficiency of word lattice decoding

WMT06 German→English Test-Set Stats

	Nodes	Length	Paths	Decoding time
Surface	(27.8)	27.8	1	43 sec/sent
Split	(31.4)	31.4	1	-
Lattice	40.7	31.4	1.7×10^9	52 sec/sent

German-English

- German
  - Fusional inflection (handful of forms)
  - Considerable productive compounding

Language	*Tokens*	*Types*	*Singletons*
German	14.6M	190k	95k
-stem	“	155k	82k
-split*	16.3M	83k	33k
-stem+split	“	67k	29k
English	15.3M	65k	24k

*** P. Koehn and K. Knight. (2003) Empirical Methods for Compound Splitting**

German-English

- What to do about the penalty function when you can split compounds and stem?

Er gab uns Übungsblätter	(surface)
Er gab uns Übungsblatt	(stem)
Er gab uns Übung Blätter	(split)
Er gab uns Übung Blatt	(stem+split)

- Ideally, two features (weighted or binary): one for splitting and the other for stemming

Results for Word Lattices

- Europarl German→English
  (WMT06 Shared Task, same as Y&K)

	BLEU*
Surface-only	25.55
Lattice (surface-only training)	25.70
Lattice (combined models)	25.69

*** 1 reference translation**

Arabic-English

- Arabic segmentation / tokenization / normalization is commonly reported to help (but this is not uncontroversial)

  alra'iis → *al ra'iis*

  sayusaafaru → *sawfa yusaafaru*

- Does segmentation help? Does it lose some important information?
  - Use word lattices to find out!

Results for Word lattices

- GALE MT03 Arabic → English

Input	BLEU*
Unsegmented	48.12
Segmented	49.20
Seg+Noseg (Lattice)	**49.70**

*** 4 reference translations**

Conclusion

- Word lattices and CNs have applications aside from speech recognition.
- Preprocessing decisions, such as backoff, can sometimes be better made by the decoder (cf. Czech-English results)
- How much of a problem is morphological sparseness?

Thank You!

Acknowledgements:

Nicola Bertoldi
David Chiang
Marcello Federico
Philipp Koehn
Adam Lopez
Philip Resnik
Daniel Zeman
Richard Zens