



Text-to-Text Generation

Katja Filippova
katjaf@google.com

Google research

This course

- A quick overview of a number of topics under the umbrella term “text-to-text generation”.
 - Research problems - what is being done and why?
 - Common approaches - how are the problems tackled? [Intuition, not an in-depth presentation! Lot of handwaving!]
 - Pointers to related literature: where to read more?
 - Pointers to useful data: how to try stuff out?
- Ambition: get you interested in learning more and doing research on those topics (see papers coming from Russia at [NE]ACL, EMNLP, Coling, in 2012 - ...).



Generation



- “Is the natural language processing task of generating natural language from a machine representation such as a knowledge base or a logical form” (from Wikipedia).
- Sometimes is seen as a counterpart to natural language understanding (esp. syntactic and semantic parsing).
- SIGGEN = Special Interest Group in GENeration.
- By far smaller community than the NLU one.



NLG from logical forms

$\lambda x.state(x) \wedge loc(miss_river, x)$

NLG from logical forms

$\lambda x.state(x) \wedge loc(miss_river, x)$

Which states does the Mississippi run through?



Data-to-text generation

- Weather forecast (temperature, rain likelihood, wind).



Data-to-text generation

- Weather forecast (temperature, rain likelihood, wind).



Data-to-text generation

- Sports competitions (scores, teams, players).

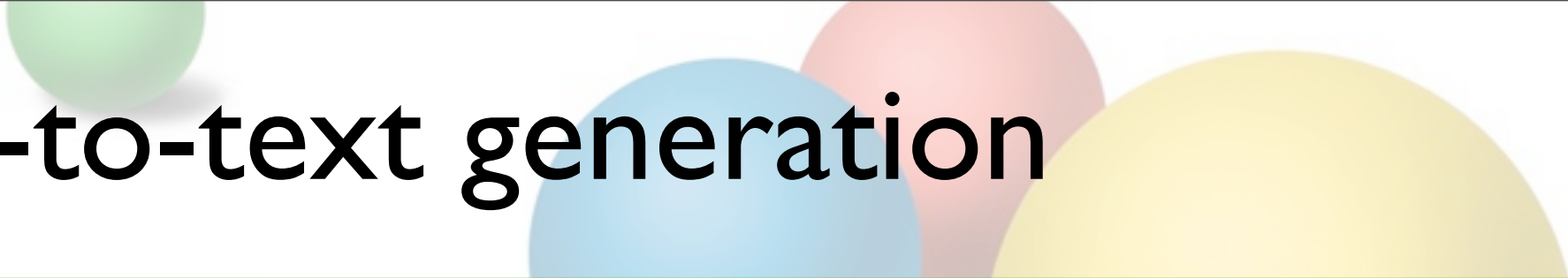
1. German League 2006/2007 - Table - Simple View

>1. German League | 2. German League | 3. German League - South | 3. German League - North

>Simple View | Advanced view | Results Grid | Fixtures | Results

2001/02 | 2002/03 | 2003/04 | 2004/05 | 2005/06 | >2006/07 | 2007/08 | 2006/07

NO	NAME	PLD	W	D	L	GF	GA	PTS
1	Stuttgart	34	21	7	6	61	37	70
2	Schalke 04	34	21	5	8	53	32	68
3	Bremen	34	20	6	8	76	40	66
4	Bayern München	34	18	6	10	55	40	60
5	Leverkusen	34	15	6	13	54	49	51
6	Nürnberg FC	34	11	15	8	43	32	48
7	Hamburg SV	34	10	15	9	43	37	45
8	Bochum	34	13	6	15	49	50	45
9	Dortmund	34	12	8	14	41	43	44
10	Hertha Berlin	34	12	8	14	50	55	44
11	Hannover 96	34	12	8	14	41	50	44
12	Bielefeld	34	11	9	14	47	49	42



-to-text generation

- Route instructions (map, streets, landmarks).

А

Россия, Санкт-Петербург, Университетс

✕

Б

Эрмитажный мост

✕

Проложить

Найдено 2 варианта проезда:

16 мин.

без пересадок, пешком ≈ 1,1 км

16 мин.

без пересадок, пешком ≈ 700 м

🚶

А

Университетская набережная

⋮

Пешком (1,1 км по прямой)

≈ 16 мин.

Б

Эрмитажный мост

Data-to-text generation

- Standard pipeline:
 - content selection (what to say);
 - document/sentence planning (where to say what, aggregation);
 - surface realization (lexical choice, referring expression generation, syntax, morphology, word order).



D2T subtasks

- GRE (= generating referring expressions).



D2T challenges

- GIVE (= generating instructions in virtual environments)



Why T2T?

- Tons of information in text format (news, blogs, reviews, ...) which we would like to understand and use.
- Major application - **text summarization**:
“the creation of a much shorter text from a collection of related documents which contains the most important points from the input”.
 - What is “most important”?
 - generic importance, or
 - interesting for this user, or
 - related to the query, or ...



Why T2T?

- Question/answer generation:

Валентина Ивановна Матвиенко (в девичестве Тютина) родилась 7 апреля 1949 года в городе Шепетовка Хмельницкой области Украинской ССР.

- *Где родилась губернатор Санкт-Петербурга?*
- *В городе Шепетовка.*

Klaus Wowereit wurde am 1. Oktober 1953 als jüngstes von fünf Kindern in Berlin geboren.

- *Wo wurde der Bürgermeister von Berlin geboren?*
- *In Berlin.*



Why T2T?

- Text simplification (make understandable to children/non-native speakers).

Whilst the process of meiosis bears a number of similarities with the the 'life-cycle' cell division process of **mitosis**, it differs in two important respects:-

- the chromosomes in meiosis undergo a recombination which shuffles the genes producing a different genetic combination in each gamete, compared with the co-existence of each of the two separate pairs of each chromosome (one received from each parent) in each cell which results from mitosis.

Meiosis is a special type of **cell division**. Unlike **mitosis**, the way normal body cells divide, meiosis results in cells that only have half the usual number of chromosomes, one from each pair. For that reason, meiosis is often called *reduction division*. In the long run, meiosis increases **genetic** variation, in a way which will be explained later.



Why T2T?



- Q1: What other text-to-text applications you can think of? Provide a motivating example.



T2T subtasks

- Sentence compression (remove unimportant content from summary sentences).
- Sentence fusion (combine several sentences).
- Paraphrasing (find a better wording while keeping the meaning).
- Sentence ordering (make summary coherent).



i. Paraphrasing

- “A *paraphrase* is an alternative surface form in the same language expressing the same semantic content as the original form” (Madnani&Dorr, CL’10).
- **lexical** paraphrases: synonyms / hypernyms from WordNet, eg, *car - automobile, месяц - луна, eat - devour*.
- **phrasal** paraphrases: *X bought Y from Z - Z sold Y to X, X invented Y - X is the inventor of Y - Y is an invention of X*.
- **sentential** paraphrases:
Harry Potter creates magic at the box office
Last Harry Porter movie sees best opening of all time
Harry Porter finale shatters weekend record



i. Paraphrasing: Why?

- Where would one need paraphrases?
 - query and pattern expansion:
‘ways to live with feline allergy’ - ‘how to deal with cat allergens’.
 - machine translation: *‘Sie war schon Wurzel’* (R.M.Rilke)
и превратилась в корень
успела стать она подземным корнем
она была лишь корнем
была она лишь корнем
была она уже подобна корню
была она как корень
 - summarization: same but with shorter words.



i. Paraphrasing: How?

- Data-driven approaches: paraphrasing with corpora.
 - Huge, single corpus.
 - Monolingual parallel corpus.
 - Monolingual comparable corpus.
 - Bilingual parallel corpus.



i. Paraphrasing: How?

- Single but huge monolingual corpus.
- Distributional similarity: words/phrases appearing in similar contexts must be somehow similar.
 - things that can *be big, red, heavy, small, dark, interesting, boring.*
 - things that can *lie on the desk, bed, table, chair, shelf.*
 - things we can *buy in a shop, kiosk or bookstore, lend to a friend, forget on the train, win the Nobel prize for, write in the early XIX century, publish at MIT press, get per post.*



i. Paraphrasing: How?

- Single but huge monolingual corpus.
- Distributional similarity: words/phrases appearing in similar contexts must be somehow similar.
 - things that can *be big, red, heavy, small, dark, interesting, boring.*
 - things that can *lie on the desk, bed, table, chair, shelf.*
 - things we can *buy in a shop, kiosk or bookstore, lend to a friend, forget on the train, win the Nobel prize for, write in the early XIX century, publish at MIT press, get per post.*

Q2: Why should two words be similar if they share many contexts?



i. Paraphrasing: How?

- Pasca & Dienes 2005, web-scale corpus:
 - Extract ngrams with the minimum length from the corpus.
 - Break every ngram into 'left-ctxt : candidate : right-ctxt', eg,
Synthetic drug law became effective this week.
Synthetic drug law came into effect recently.
Synthetic drug law went into effect this month.
 - Measure candidate similarity by counting overlap in contexts.
 - The method works on word sequences, no structural information.

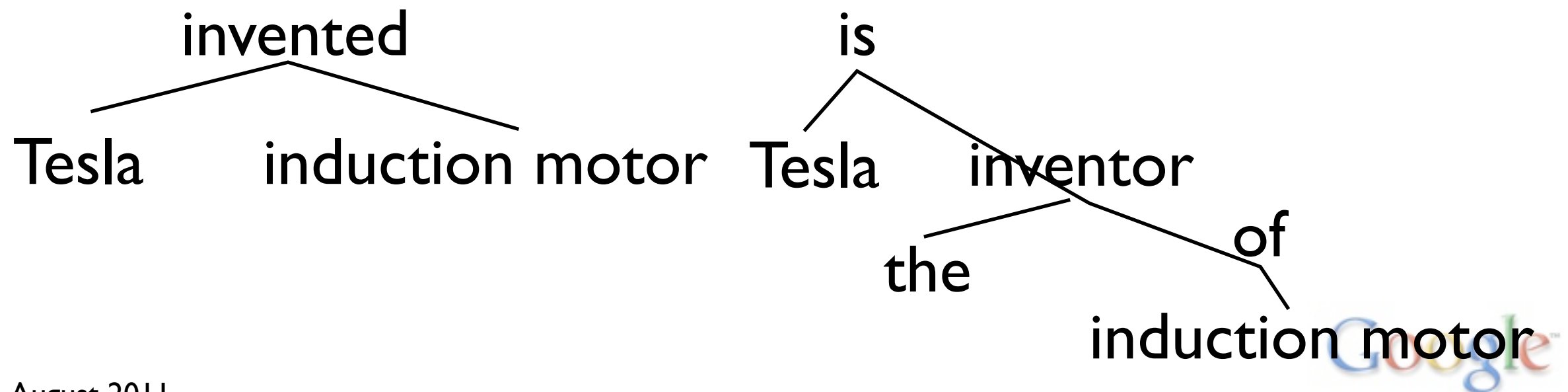


i. Paraphrasing: How?

- Pasca & Dienes 2005, web-scale corpus:
 - Extract ngrams with the minimum length from the corpus.
 - Break every ngram into 'left-ctxt : candidate : right-ctxt', eg,
Synthetic drug law became effective *this week*.
Synthetic drug law came into effect *recently*.
Synthetic drug law went into effect *this month*.
 - Measure candidate similarity by counting overlap in contexts.
 - The method works on word sequences, no structural information.

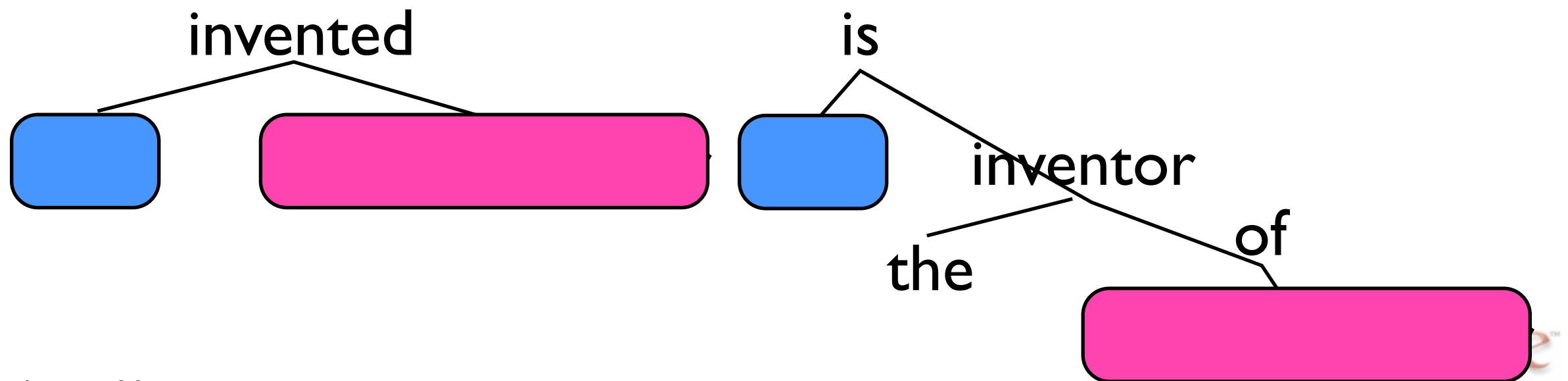
i. Paraphrasing: How?

- Lin & Pantel, 2001:
 - Structural representation: dependency trees.
 - Extract generalized paraphrase templates from **dependency paths**:
X invented Y = X is the inventor of Y.
 - If two dependency paths tend to link the same words, they are likely to be paraphrases - the same idea of distributional similarity.



i. Paraphrasing: How?

- Lin & Pantel, 2001:
 - Structural representation: dependency trees.
 - Extract generalized paraphrase templates from **dependency paths**:
X invented Y = X is the inventor of Y.
 - If two dependency paths tend to link the same words, they are likely to be paraphrases - the same idea of distributional similarity.



i. Paraphrasing: How?

- Lin & Pantel 2001:
 - Path similarity:
 $\text{sim}(p, p') = \sqrt{\text{sim}(X, X') \times \text{sim}(Y, Y')}$
 - Slot similarity w.r.t. p & p' - $\text{sim}(X, X')$ - looks at how many words appear in *both* slots relative to the number of words appearing in *any* of the two slots.
 - Words are not equally weighted: 'he' has less weight than 'Barack Obama' and is a weaker signal of path similarity.
 - Mutual information-inspired measure of the association of word w and slot s in path p .

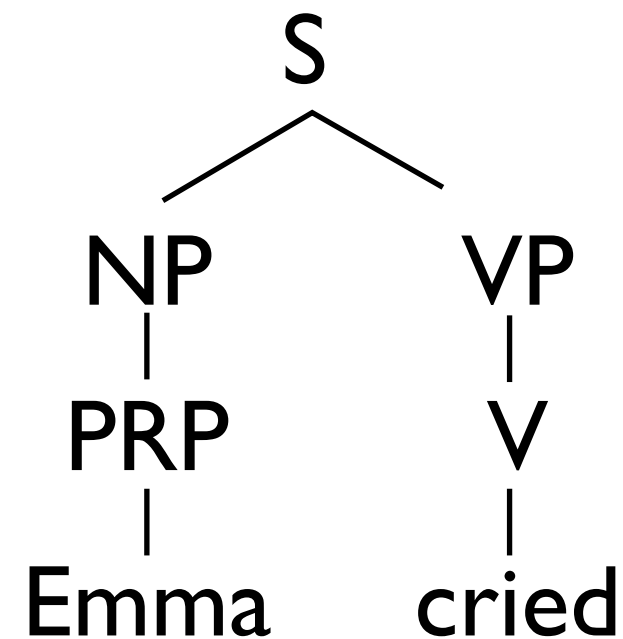
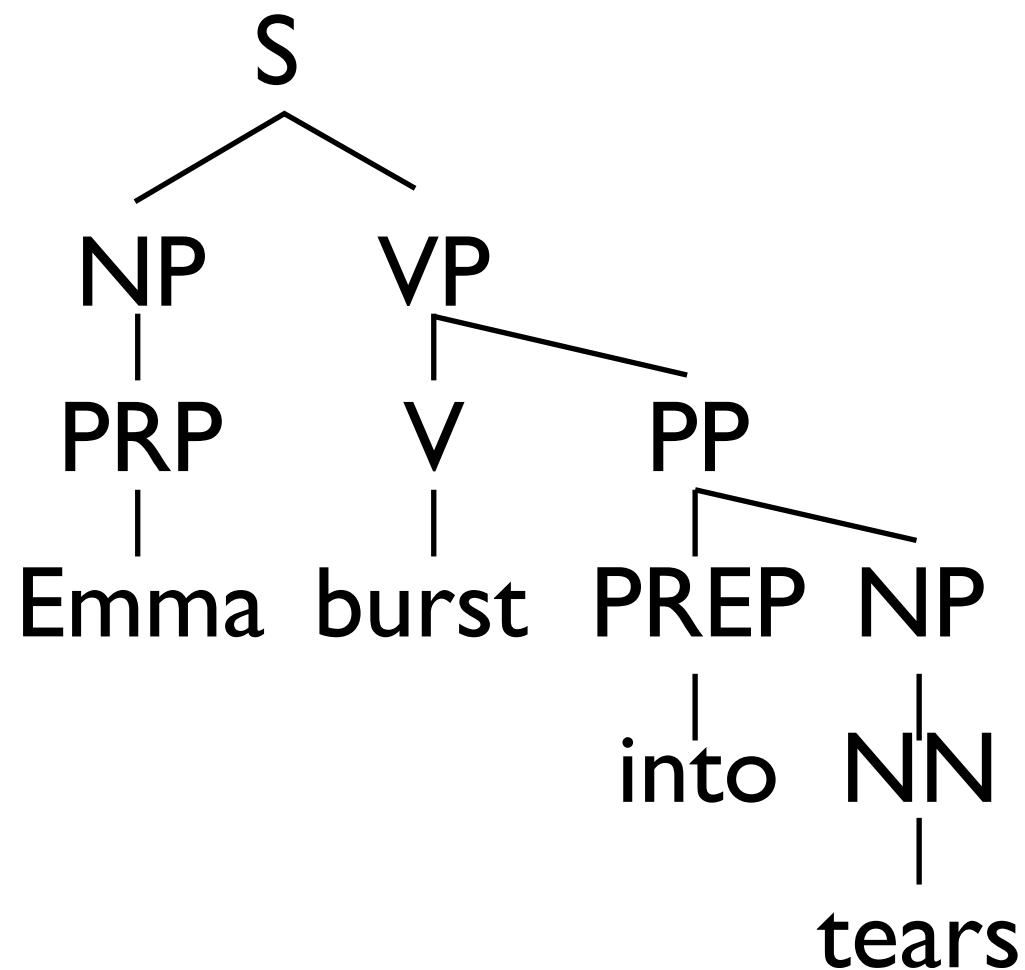
i. Paraphrasing: How?

- Monolingual parallel corpus.
- Machine learning-based approach (Barzilay & McKeown, 2001):
 - Data - multiple fiction translations:
Emma burst into tears and he tried to comfort her.
Emma cried and he tried to console her. (“Madame Bovary”)
 - Extract pairs which are positive+ (*<he, he>*, *<tried, tried>*) and negative- (*<he, tried>*, *<Emma, console>*) examples.
 - For every pair, extract contextual features.
 - Feature strength is the MLE - $|f|_+ / (|f|_+ + |f|_-)$.
 - Find more paraphrases, update weights, repeat.



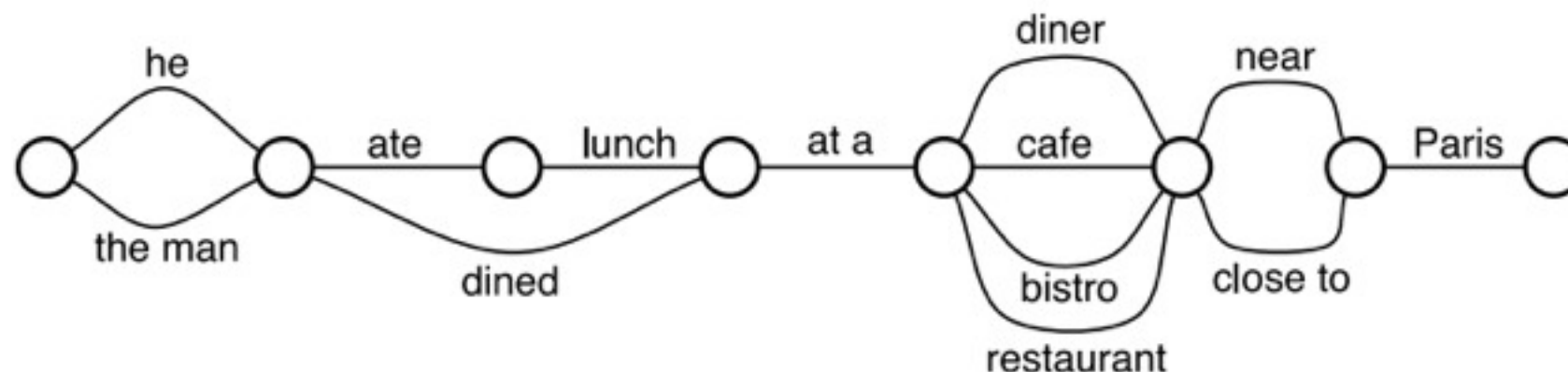
i. Paraphrasing: How?

- Pang, Knight & Marcu 2003:
 - Align constituency trees of parallel sentences.



i. Paraphrasing: How?

- Quirk, Brockett & Dolan 2004:
 - Use the standard SMT formula,
 $E^* = \arg \max p(E^* | E) = \arg \max p(E^*) p(E | E^*)$
 - 140K “parallel” sentences obtained from online news (articles about the same event, edit distance to discard sentences which cannot be paraphrases).
 - Paraphrasal pairs are extracted with associated probabilities.
 - Given a sentence, a lattice of possible paraphrases is constructed and dynamic programming is used to find the best scoring paraphrase.



i. Paraphrasing: How?

- Parallel corpora are rare, comparable corpora are abundant.
- Shinyama et al. 2002:
 - News articles from two sources which appeared on the same day.
 - Similar articles are paired.
 - Preprocessing: dependency parse trees, NE recognition.
 - NEs are replaced with generic slots.
 - Patterns pointing to the same NEs are taken as paraphrases.

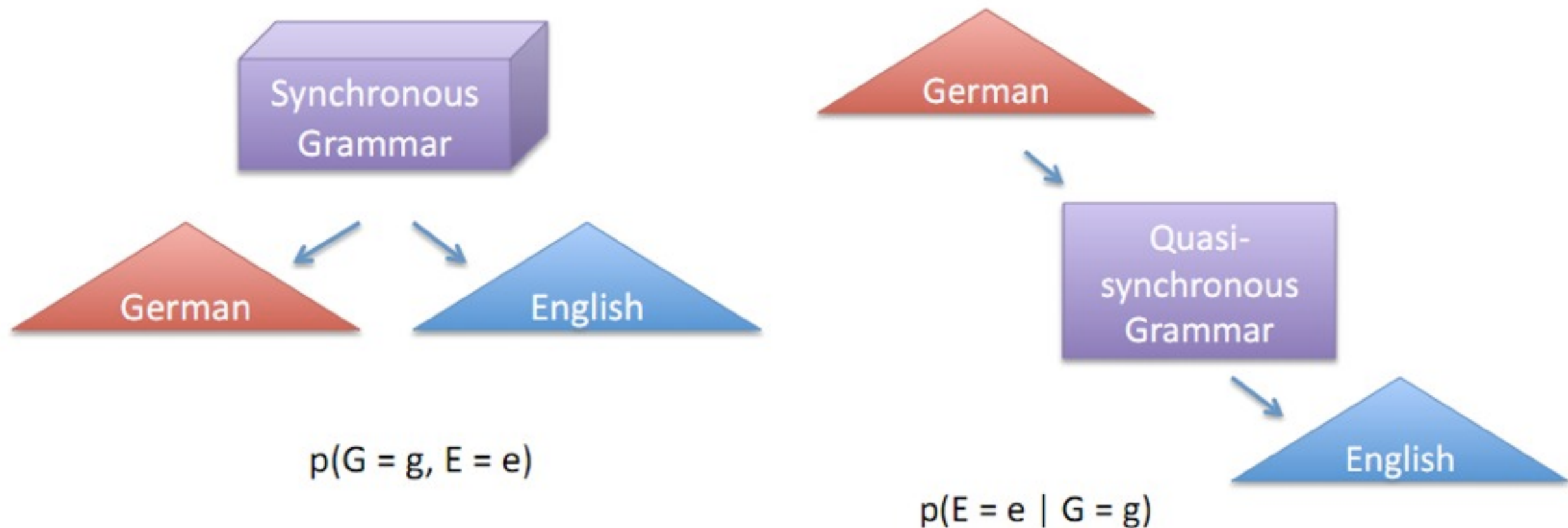
i. Paraphrasing: How?

- Barzilay & Lee 2003:
 - Two news agencies, the same period of time.
 - Similar sentences (sharing many ngrams) are clustered.
 - Multiple sequence alignment which results into a slotted word lattice.
 - Backbone nodes are identified (shared by $>50\%$ of sentences) as points of commonality.
 - Variability signals argument slots.
 - Given a new sentence, a suitable cluster needs to be found before a paraphrase can be generated (there might be no such cluster).



i. Paraphrasing: How?

- Use of Synchronous and Quasi-synchronous Grammars.



(these pictures are stolen from the presentation of Noah Smith at T2T workshop, ACL'11)

i. Paraphrasing: How?

- Synchronous grammars:
 - define pairs of rules, e.g., for German and English:
(VP; VP) \rightarrow (V NP; NP V)
 - can be probabilistic (compare with PCFGs).
 - does not have to be constituency syntax, e.g., TAG and logical forms (Shieber & Shabes, 1990).
 - have been used for MT and also for getting paraphrase grammars.

i. Paraphrasing: How?

- Quasi-synchronous grammars (Smith & Eisner, 2006):
 - were introduced for MT.
 - the output sentence is “inspired” by the source sentence, not determined.
 - again, does not have to be constituency syntax, e.g., dependency representation.
 - have been used for other text-to-text generation tasks, like text simplification (Woodsend & Lapata, 2011) or question generation (Wang et al. 2007).



i. Paraphrasing

Questions?



ii. Sentence compression

- Simple and intuitive idea which is to shorten a long sentence preserving the main points and removing less relevant information.

ii. Sentence compression

- Simple and intuitive idea which is to shorten a long sentence preserving the main points and removing less relevant information.

Simple and intuitive idea which is to shorten a long sentence preserving the main points and removing less relevant information.



ii. Sentence compression

- Simple and intuitive idea which is to shorten a long sentence preserving the main points and removing less relevant information.

Simple and intuitive idea which is to shorten a long sentence preserving the main points and removing less relevant information.



ii. Sentence compression

- Simple and intuitive idea which is to shorten a long sentence preserving the main points and removing less relevant information.

Simple and intuitive idea which is to shorten a long sentence preserving the main points and removing less relevant information.

deletion

(substitution reordering)



ii. Sentence compression

- **Rule-based** approaches rely on PoS annotations and syntactic structures and remove constituents/dependencies likely to be less important (Grefenstette 1998, Corston-Oliver&Dolan 1999):
 - relative clauses, prepositional phrases
 - proper nouns > common nouns > adjectives
- Further sources of information can be used, e.g., a subcategorization lexicon (Jing 2000):

give(Subj, AccObj, DatObj)

On Friday, Ann gave Bill a book.



ii. Sentence compression

- Rules can be induced from a corpus of compressions (Dorr et al. 2003, Gagnon & DaSilva 2005):
 - what kind of PPs are removed,
 - what are the PoS, syntactic features of the removed constituents,
 - look at a manually crafted corpus or at a corpus of news headlines (compare the length of headlines with the average sentence length).
- **Supervised** approaches learn what is “removable” without direct human intervention.

ii. Sentence compression

- Knight & Marcu 2002 use the **noisy-channel model**:
 - Bayes rule: $p(y|x) = p(x,y)/p(x) = p(x|y)p(y)/p(x)$
 - Look for y maximizing $p(y|x) \sim p(x|y)p(y)$

$$\mathbf{y} = \mathbf{arg\ max\ } p(\mathbf{x}|\mathbf{y})\ p(\mathbf{y})$$

ii. Sentence compression

- Knight & Marcu 2002 use the **noisy-channel model**:
 - Bayes rule: $p(y|x) = p(x,y)/p(x) = p(x|y)p(y)/p(x)$
 - Look for y maximizing $p(y|x) \sim p(x|y)p(y)$

$$y = \arg \max p(x|y) p(y)$$

$$\text{MT: } f = \arg \max p(e|f) p(f)$$

ii. Sentence compression

- Knight & Marcu 2002 use the **noisy-channel model**:
 - Bayes rule: $p(y|x) = p(x,y)/p(x) = p(x|y)p(y)/p(x)$
 - Look for y maximizing $p(y|x) \sim p(x|y)p(y)$

$$y = \arg \max p(x|y) p(y)$$

$$\text{MT: } f = \arg \max p(e|f) p(f)$$

*Q3: Why “split”
into two things?*



ii. Sentence compression

- Knight & Marcu 2002 use the **noisy-channel model**:
 - Bayes rule: $p(y|x) = p(x,y)/p(x) = p(x|y)p(y)/p(x)$
 - Look for y maximizing $p(y|x) \sim p(x|y)p(y)$

$$y = \arg \max p(x|y) p(y)$$

$$\text{MT: } f = \arg \max p(e|f) p(f)$$

*Q3: Why “split”
into two things?*

$$\text{SC: } s = \arg \max p(s|l) p(s)$$



ii. Sentence compression

- What is $p(s)$ supposed to do?
 - assign low probability to ungrammatical, “strange” sentences.
- How to estimate $p(s)$? E.g., with a n -gram model from a corpus of (compressed) sentences.
- What is $p(l|s)$ supposed to do?
 - assign low probability to compressions which have little to do with the input,
 - assign very low probability to compressions which flip the meaning (e.g., delete *not*).
- How to estimate $p(l|s)$?



ii. Sentence compression

- Knight & Marcu 2002 look at constituency trees (CFG):

$s = S (NP (John)$
 $VP (VB (saw)$
 $NP (Mary)))$

$\sim p(s) = p(S-NP VP|S) p(NP-John|NP)$
 $p(VP-VB NP|VP) p(VB-saw|VB) p(NP-Mary|NP)$
 $p(John|eos) p(saw|John) p(Mary|saw) p(eos|Mary)$

*Q4: How can these probabilities be
acquired?*



ii. Sentence compression

- Given a corpus (Ziff-Davis) K&M want to learn probabilities of the expansion rules.
 - parse the long and the short sentence,
 - align the parse trees (not always possible, the model cannot deal with that problem),
 - do maximum likelihood estimation of rules like the following:

$$p(\text{VP-VB NP PP} \mid \text{VP-VB NP})$$

Q5: What does this rule express?

- Only 1.8% of the data can be used because the model assumes that the compressions are subsequences of the original sentences.



ii. Sentence compression

- (K&M contd.) Recall: $s = \arg \max p(s|l) p(s)$
 - for every s we know how to estimate
 - $p(s)$
 - $p(s|l)$
 - the search for the best s is called decoding, not covered here.

ii. Sentence compression

- A corpus of parsed sentence pairs (long sentence / compression) can be used in other ways.
- Nguyen et al. 2004 use Support Vector Machines (SVM) and syntactic, semantic (e.g., NE type) and other features to determine the sequence of rewriting actions (shift, reduce, drop, assign type, restore).

[Similar to the shift-reduce parsing approach of Nivre, 2003+.]



ii. Sentence compression

- Galley & McKeown 2007 also use pairs of parsed trees but do not break down the probability into two terms.
- They look for $s = \arg \max p(s, l)$
Q6: Can you explain where this formula comes from?
- Consider all possible tree pairs for s and l , then

$$p(s, l) = \sum_{(\tau_s, \tau_l)} p(\tau_s, \tau_l)$$

- G&McK also use the synchronous grammar approach.

ii. Sentence compression

- Clarke&Lapata (2006, 2007) do not rely on labeled data at all (good news). A word deletion model.
- Constraints to ensure grammaticality:
 - “if main verb, then subject”
 - “if preposition, then its object”
- Discourse constraints (lexical chains) to promote words related to the main topic.
- They also introduced corpora (written and broadcast news) which can be used to test any system.



ii. Sentence compression

- The objective function to maximize is, essentially, a linear combination of the trigram score of the compression and the informativeness of single words.

$$\max z = \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^n x_{ijk} \cdot P(w_k | w_i, w_j) + \sum_{i=1}^n y_i \cdot I(w_i)$$

- x_{ijk} represents a trigram, y_i represents a single word.
- This objective function is subject to a variety of grammar and discourse constraint on the variables.
- The (approximate) solution is found with Integer Linear Programming (ILP).



ii. Sentence compression

- What is linear programming? maximizing/minimizing a linear combination of a finite number of variables which are subject to constraints.
- Binary integer programming - all variables are 0 or 1. You can think of it as a way to select from a given set given constraints on how elements in the set can be combined.

ii. Sentence compression

- An example of a grammar constraint:
 $y_i - y_j \geq 0$ if w_j modifies w_i .
- An example of a discourse constraint:
 $y_i = 1$, if w_i belongs to a lexical chain.
- Other discourse constraints are based on the Centering theory.

ii. Sentence compression

- An
y_i :
- An
y_i :
- Oth
Cer

Bad **weather** dashed hopes of attempts to halt the **flow₁** during what was seen as a lull in the **lava's** momentum. Experts say that even if the eruption stopped **today₂**, the pressure of **lava** piled up behind for six **miles₃** would bring **debris** cascading down on to the **town** anyway. Some estimate the volcano is pouring out one million tons of **debris** a **day₂**, at a **rate₁** of 15 **ft₃** per **second₂**, from a fissure that opened in mid-December.

The Italian Army **yesterday₂** detonated 400lb of dynamite 3,500 feet up Mount Etna's slopes.

ii. Sentence compression

- An y_i - Bad weather dashed hopes of attempts to halt the flow₁ during what was seen as a lull in the lava's momentum. Experts say that even
- An y_ Bad weather dashed hopes to halt the flow during what was seen as lull in lava's momentum. Experts say that even if eruption stopped, the pressure of lava piled would bring debris cascading. Some estimate volcano is pouring million tons of debris from fissure opened in mid-December. The Army yesterday detonated 400lb of dynamite.
- On Ce in mid-December.
The Italian Army yesterday₂ detonated 400lb of dynamite 3,500 feet up Mount Etna's slopes.

ii. Sentence compression

- Evaluation:
 - **intrinsic** - dependency parse score (Riezler et al. 2003): How similar are the dependency trees of the two compressions (“gold” = created by a human, and the one the system produced). The larger the overlap in dependencies, the better.
 - **extrinsic** - in the context of a QA task: Given a compressed document and a number of questions about the document, can human readers answer those questions? (The questions were generated by other humans who were given uncompressed documents.)

ii. Sentence compression

Questions?

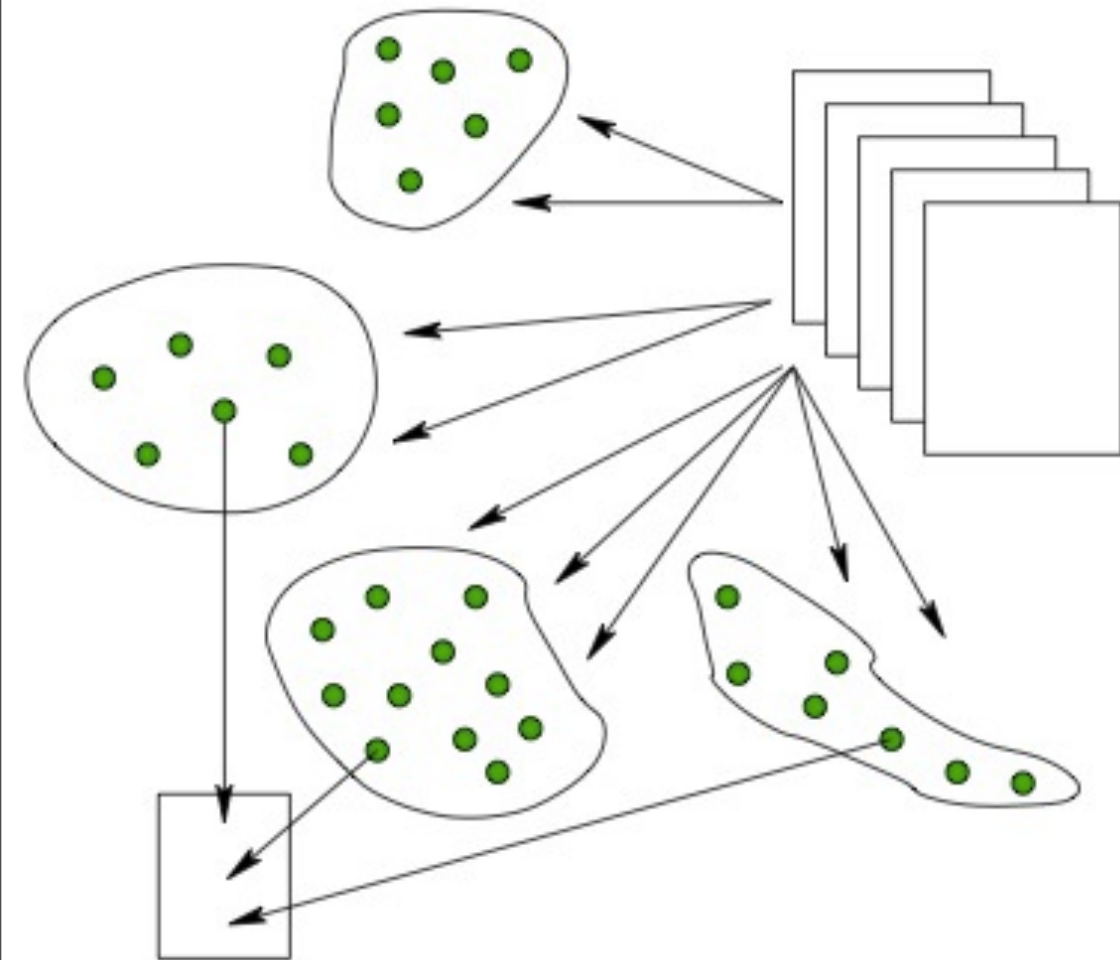


iii. Sentence fusion

- How about cases where we have several sentences as input - the multi-document summarization scenario? What can we do with them if they are somewhat similar?
- Compression is helpful if we are doing single-document summarization - we can compress every sentence we want to add to the summary, one by one.
- In case of MDS, one usually first clusters all the sentences, then ranks those clusters, then selects a sentence from each of the top N clusters.



iii. Sentence fusion



- **Extractive** approach:
 - Similar sentences are clustered.
 - Clusters are ranked.
 - A sentence is selected from each of the top clusters.

iii. Sentence fusion

- “Fuse” several related sentences into one (Barzilay & McKeown, 2005)
- Setting: multi-document, generic news summarization. Idea: recurrent information is important.
 1. *IDF Spokeswoman did not confirm this, but said the Palestinians fired an antitank missile at a bulldozer.*
 2. *The clash erupted when Palestinian militants fired machine guns and antitank missiles at a bulldozer that was building an embankment in the area to better protect Israeli forces.*
 3. *The army expressed “regret at the loss of innocent lives” but a senior commander said troops had shot in self-defense after being fired at while using bulldozers to build a new embankment at an army base in the area.*

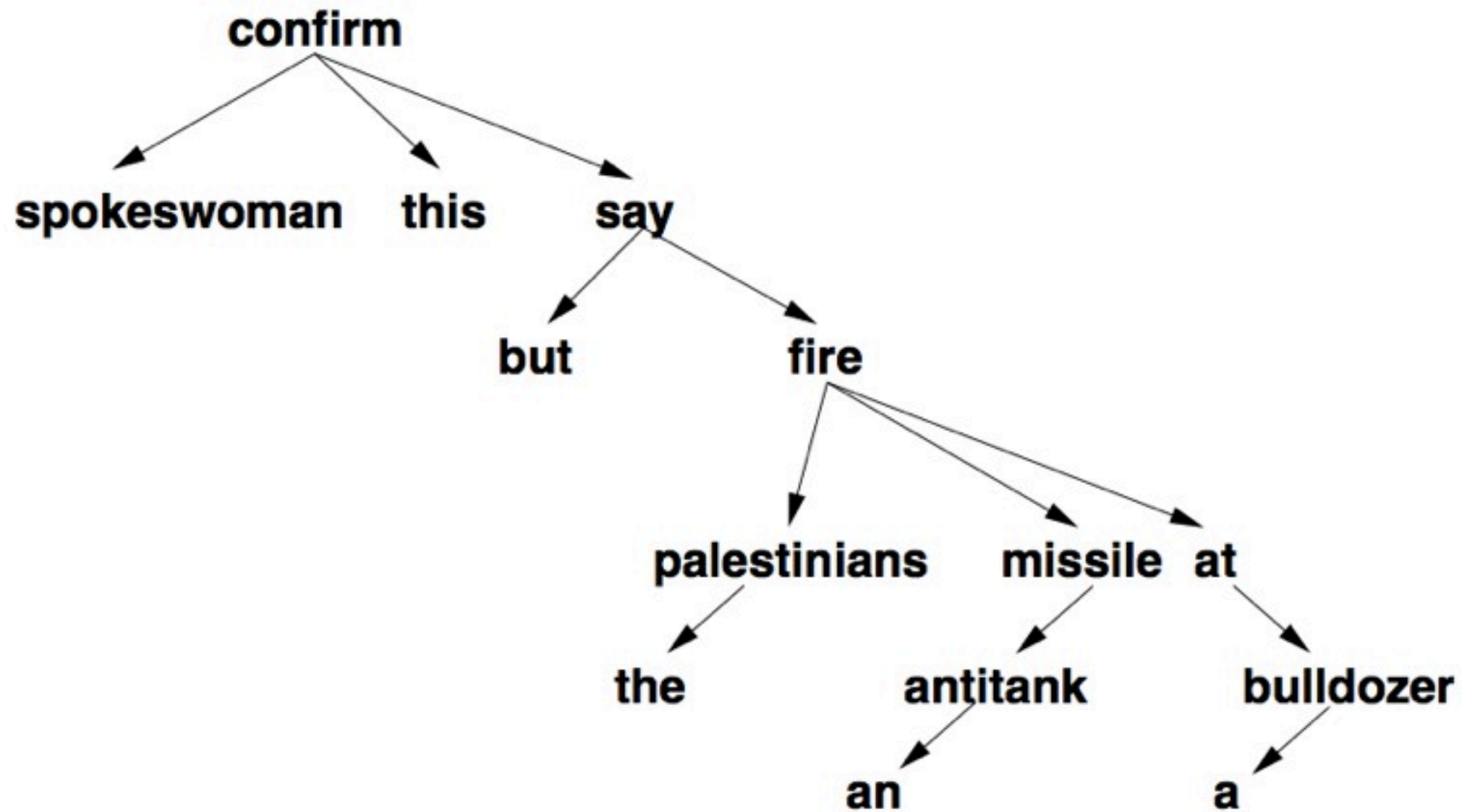


iii. Sentence fusion

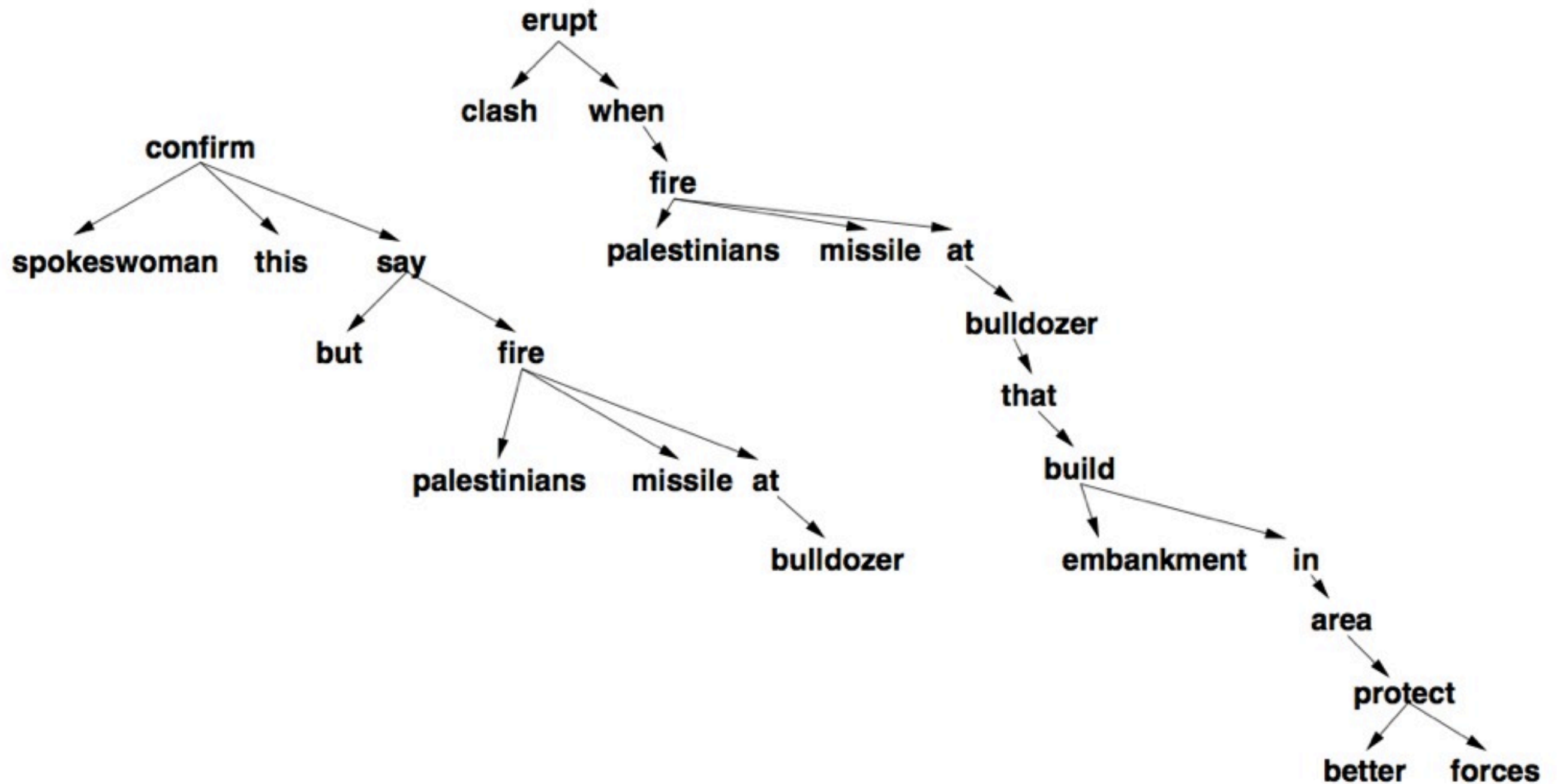
- “Fuse” several related sentences into one (Barzilay & McKeown, 2005)
- Setting: multi-document, generic news summarization. Idea: recurrent information is important.
 1. *IDF Spokeswoman did not confirm this, but said the Palestinians fired an antitank missile at a bulldozer.*
 2. *The clash erupted when Palestinian militants fired machine guns and antitank missiles at a bulldozer that was building an embankment in the area to better protect Israeli forces.*
 3. *The army expressed “regret at the loss of innocent lives” but a senior commander said troops had shot in self-defense after being fired at while using bulldozers to build a new embankment at an army base in the area.*



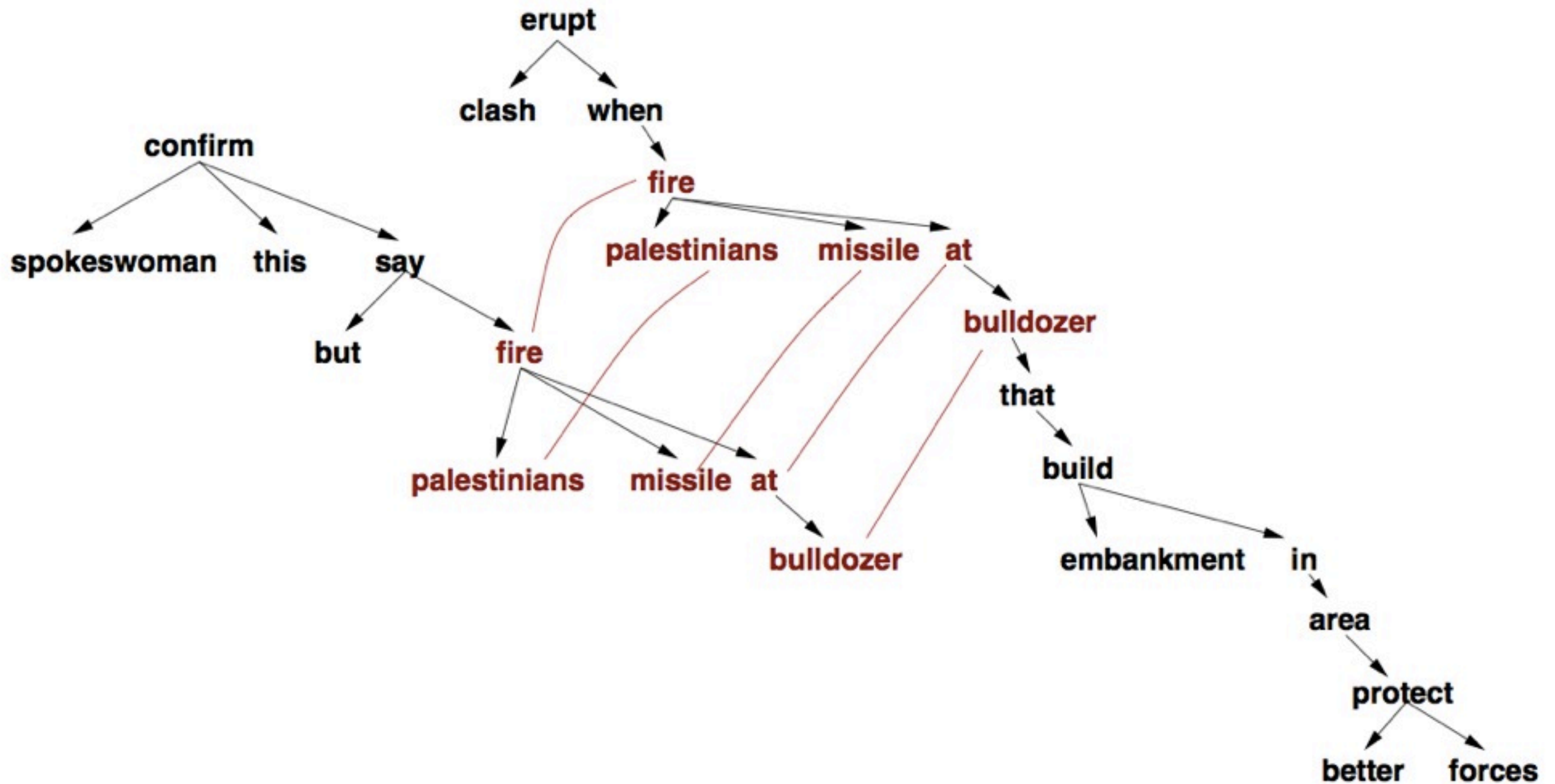
iii. Sentence fusion



iii. Sentence fusion



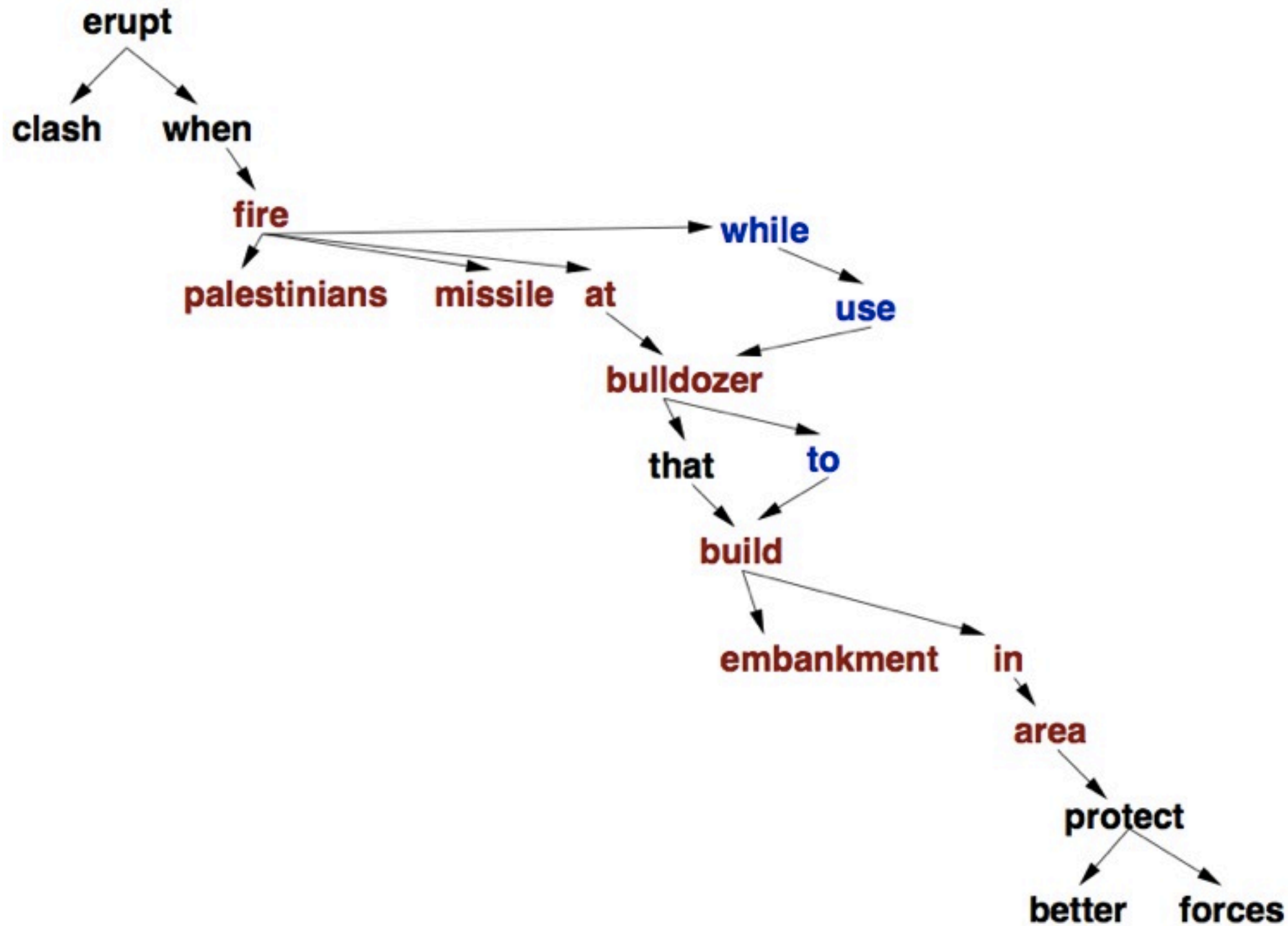
iii. Sentence fusion



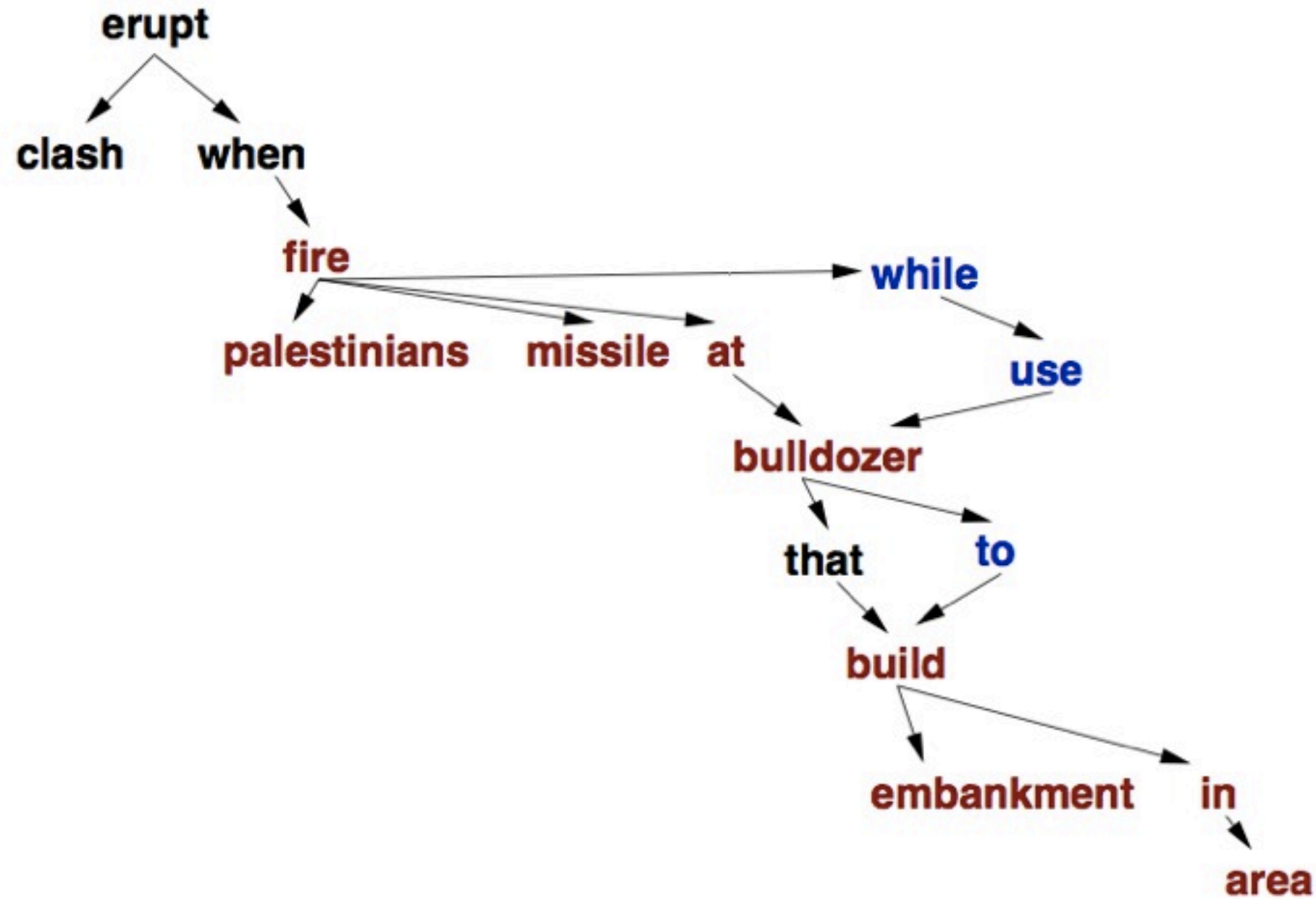
iii. Sentence fusion

- pairwise recursive bottom-up tree alignment
- each alignment has a score - the more similar two trees are, the higher the score
- from the alignment score the basis tree is determined
- it is the basis tree around which the fusion is performed

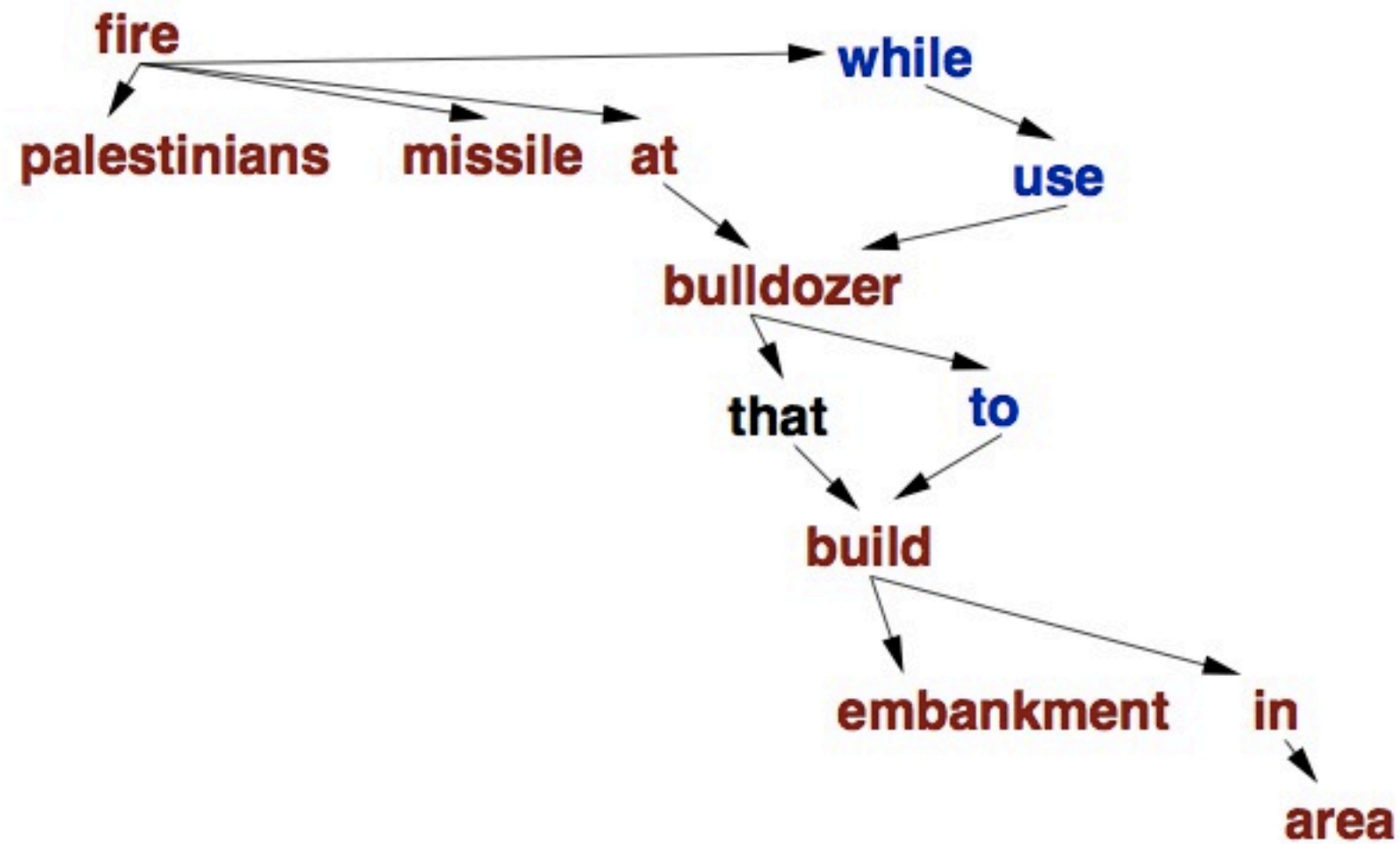
iii. Sentence fusion



iii. Sentence fusion



iii. Sentence fusion



iii. Sentence fusion

- Now we have a dependency graph expressing the recurrent content from the input.

“Overgenerate-and-rank” approach: consider up to 20K possible strings and rank them with a language model.



iii. Sentence fusion

- Now we have a dependency graph expressing the recurrent content from the input.

Q6: How can we get a sentence?

“Overgenerate-and-rank” approach: consider up to 20K possible strings and rank them with a language model.



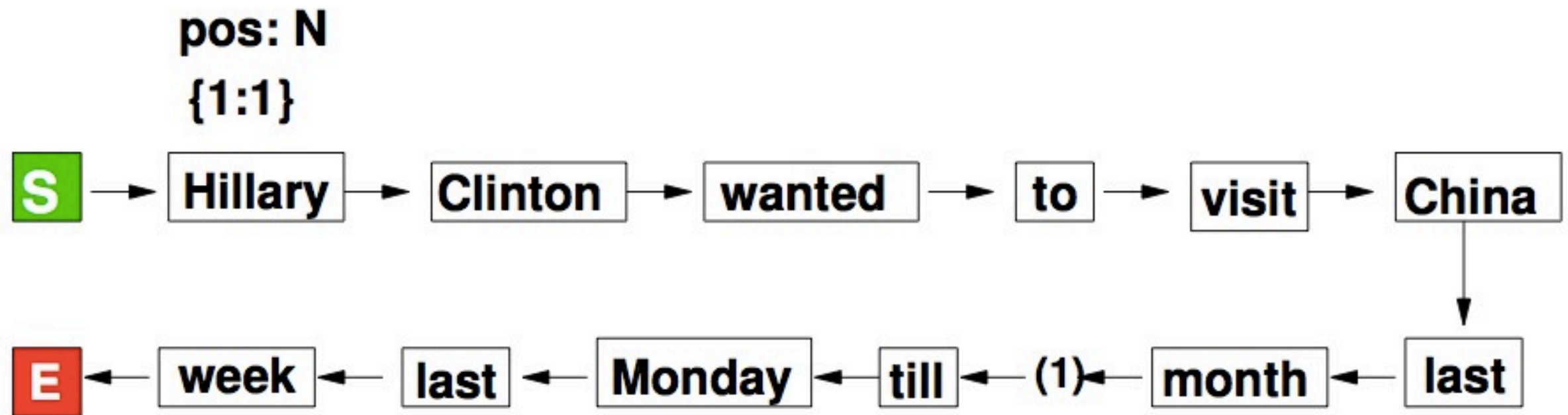
iii. Sentence fusion

- The fusion model of Barzilay & McKeown does **intersection** fusion - it relies on the idea that recurrent = important, the fused sentences express the content shared among many sentences.
- We can think of it as **multi-sentence compression**.
- Can we do without dependency representations? Let's consider a word graph where edges represent adjacency relation (Filippova 2010).

iii. Sentence fusion

- *Hillary Clinton paid a visit to the People's Republic of China on Monday.*
- *Hilary Clinton wanted to visit China last month but postponed her plans till Monday last week.*
- *The wife of a former U.S. president Bill Clinton Hillary Clinton visited China last Monday.*
- *Last week the Secretary of State Ms. Clinton visited Chinese officials.*

iii. Sentence fusion



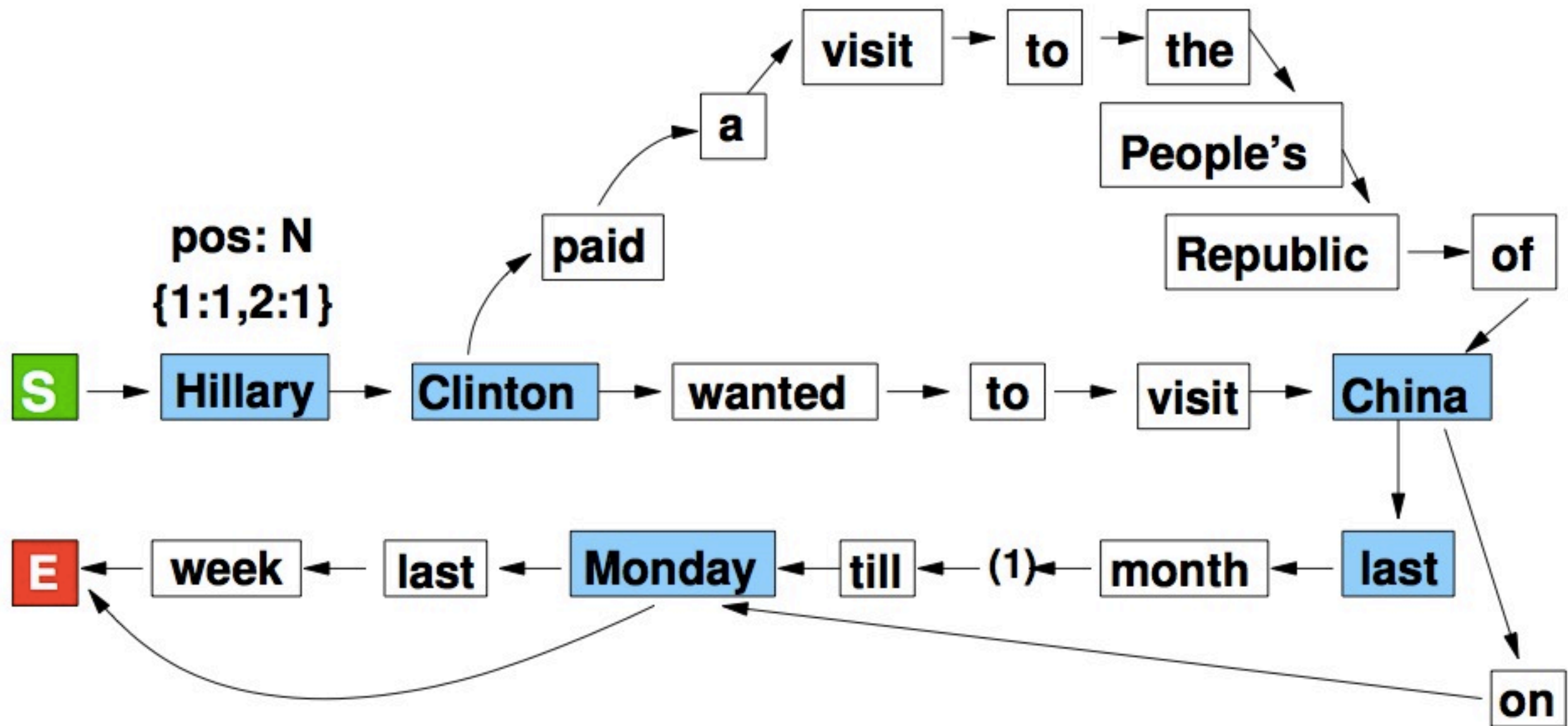
(1) but postponed her plans

iii. Sentence fusion

- Words from a new sentence are added in three steps:
 - unambiguous non-stopwords - either merged with a word-node in the graph, or a new word-node is created;
 - ambiguous non-stopwords - select the word-node with some overlap in neighbors (i.e., previous-following words in the sentence and neighbors in the graph);
 - stopwords - only merged with an existing word-node if the following word in the sentence matches an out-neighbor in the graph, otherwise a new word-node is created.
- Words from the same sentence are never merged in one node.



iii. Sentence fusion



(I) *but postponed her plans*

iii. Sentence fusion

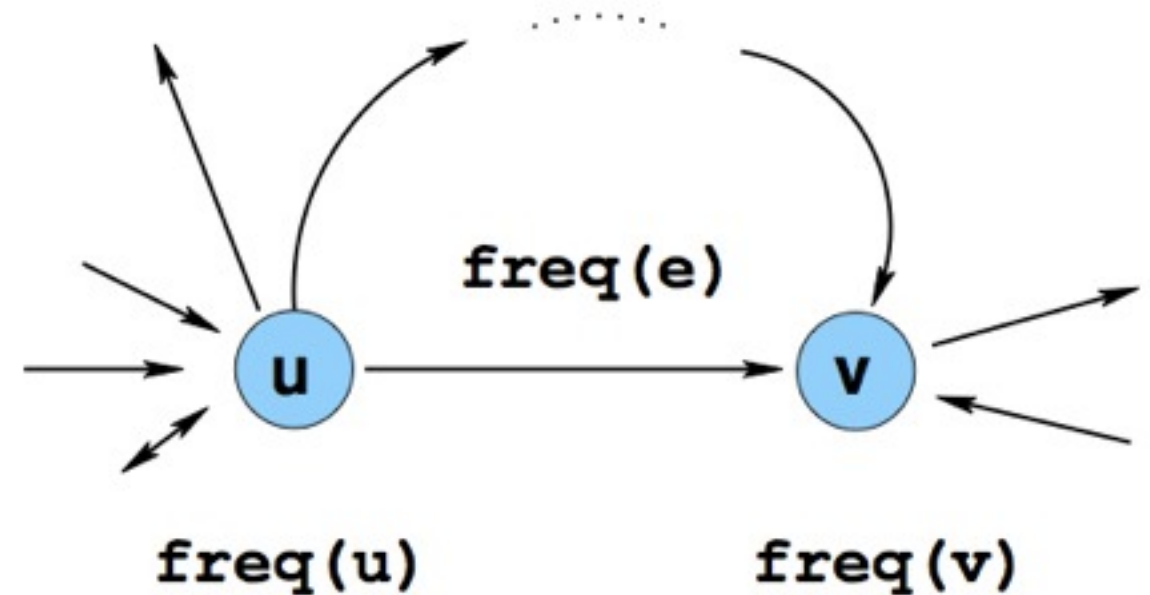
- Idea: good compressions - salient and short paths from Start to End.
- Edge weight can be defined as:

$$w(e) = \frac{1}{\text{freq}(e)}$$

$$w(e) = \frac{1}{\sum_{s \in S} \text{distance}(s, u, v)^{-1}}$$

$$w(e) = \frac{\text{freq}(u) + \text{freq}(v)}{\sum_{s \in S} \text{distance}(s, u, v)^{-1}}$$

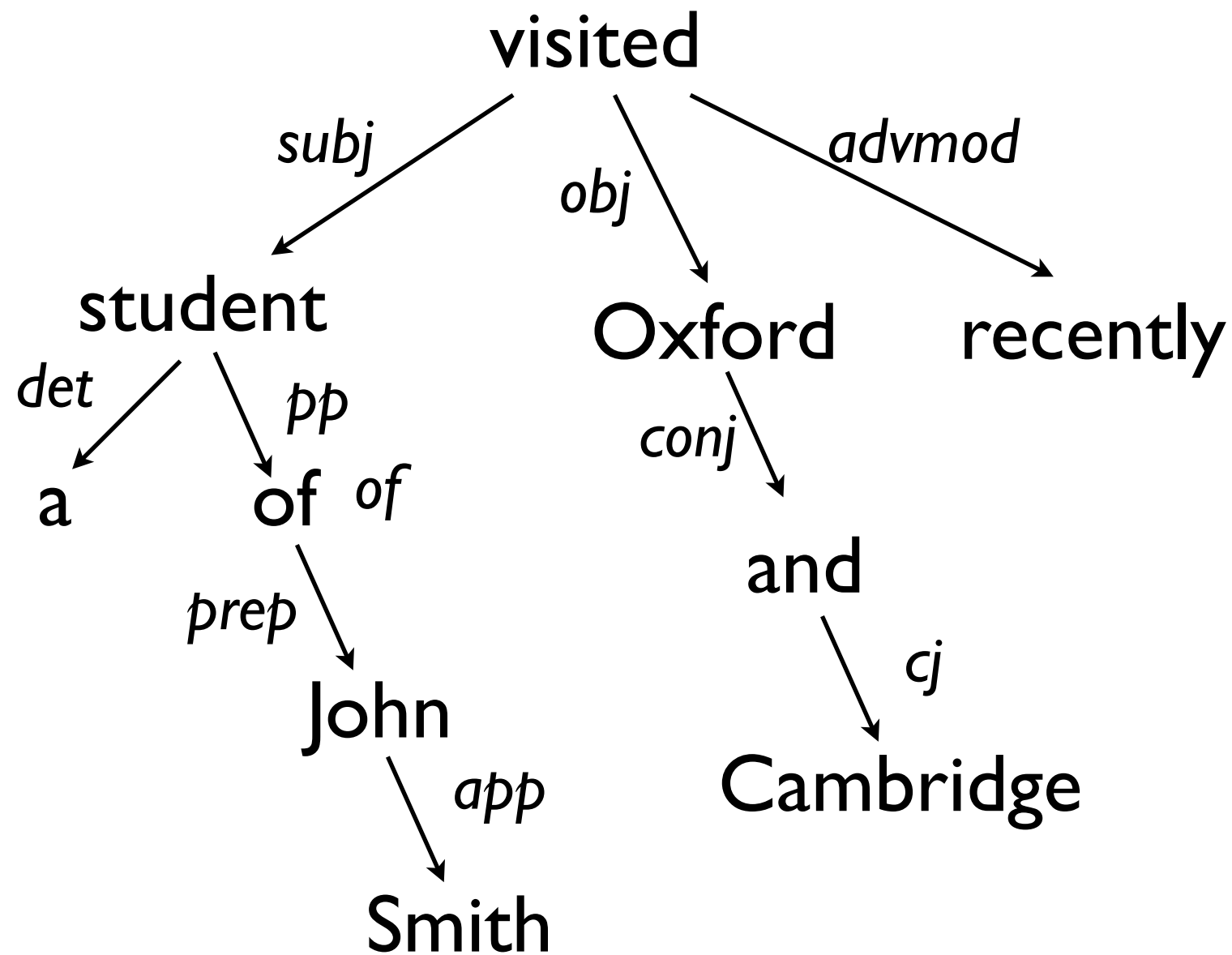
$$w(e) = \frac{\text{freq}(u) + \text{freq}(v)}{\text{freq}(u) \times \text{freq}(v) \times \sum_{s \in S} \text{distance}(s, u, v)^{-1}}$$



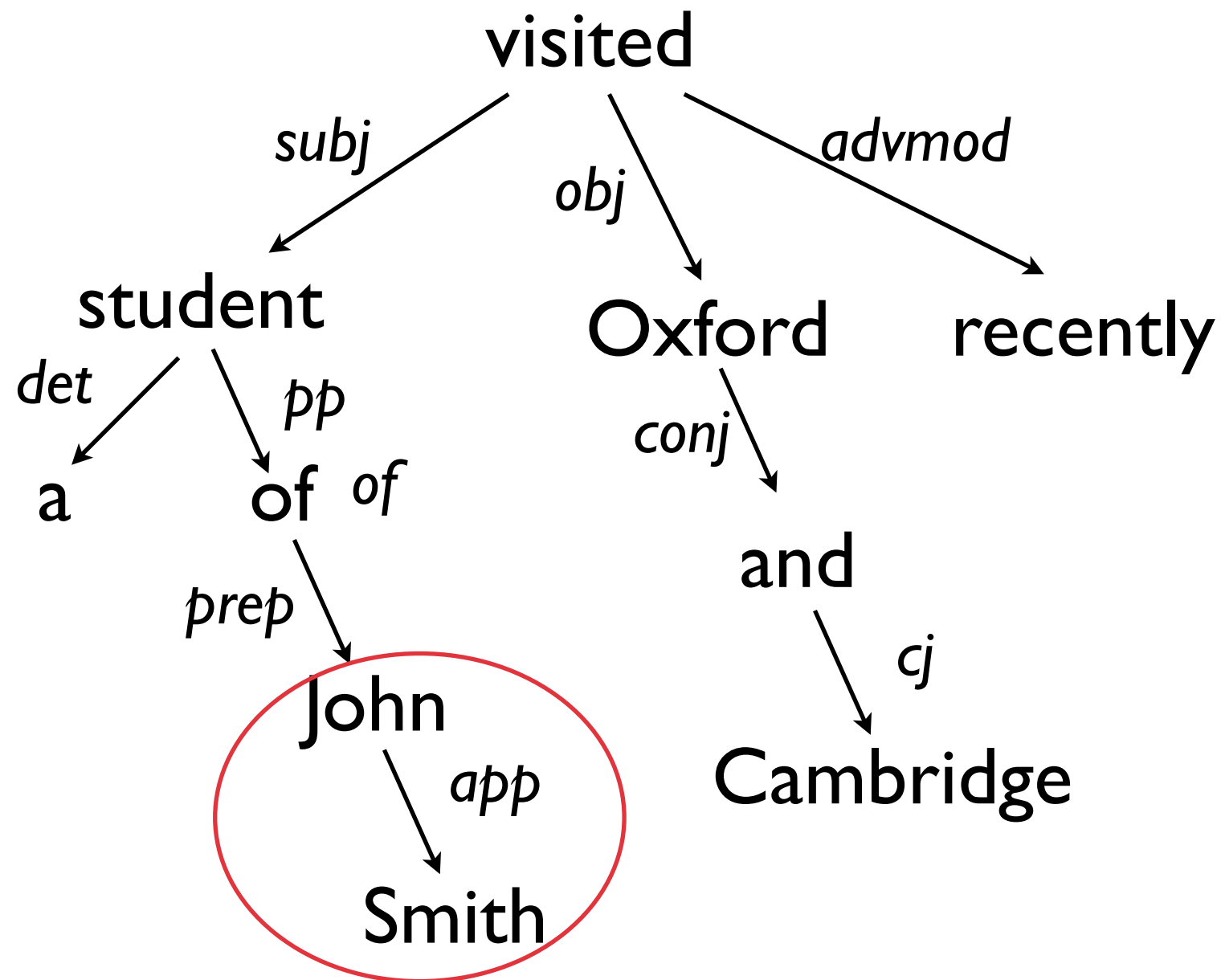
iii. Sentence fusion

- How about we are not going for the recurrent information but want to combine complementary content?
That is, we are interested not **intersection** but **union** fusion (Krahmer & Marsi, 2008).
- Can we abstract to a non-redundant representation of all the content expressed in the input? (which is a set of related sentences).
- First, can we make the dependency representation a bit more semantic?

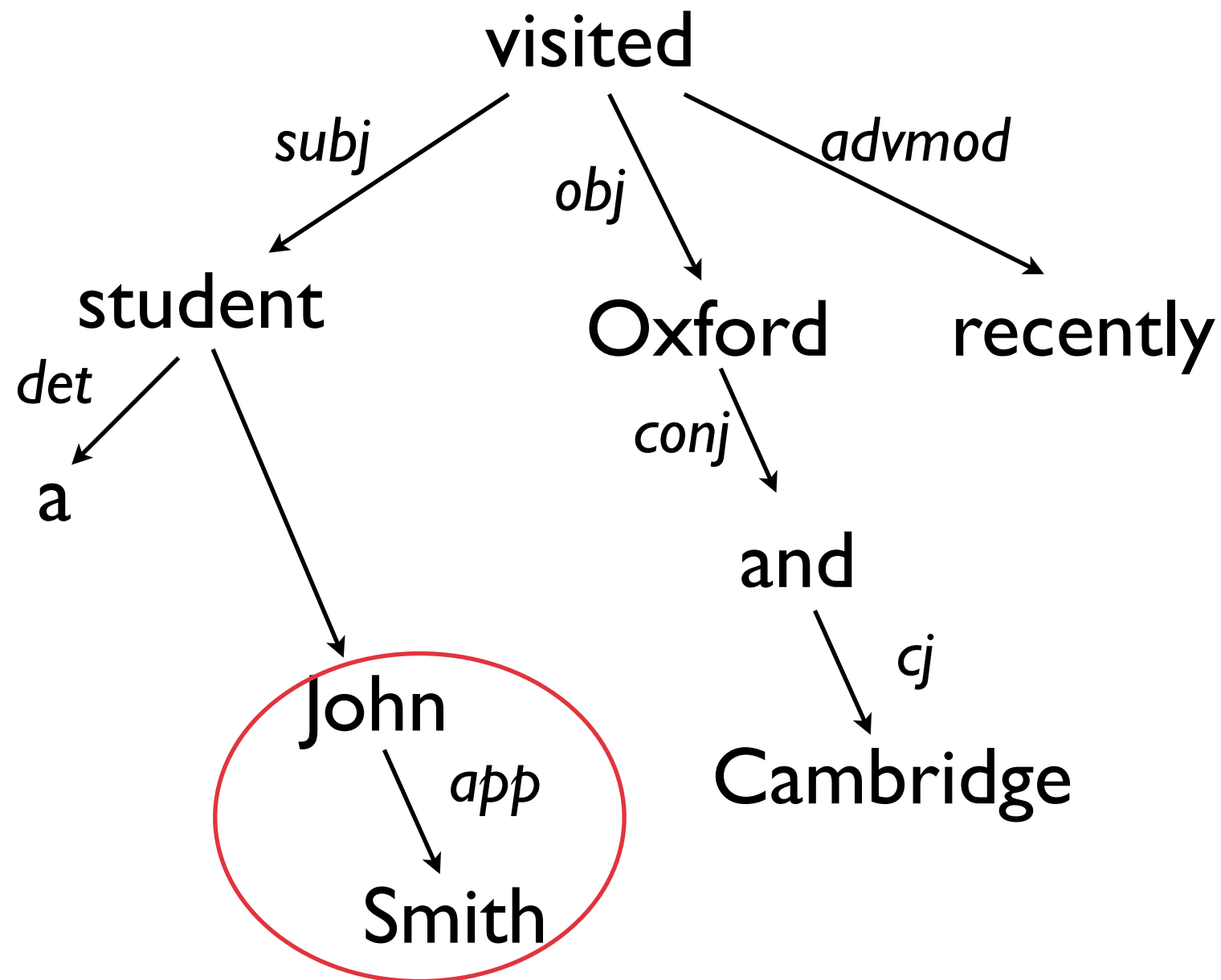
iii. Sentence fusion



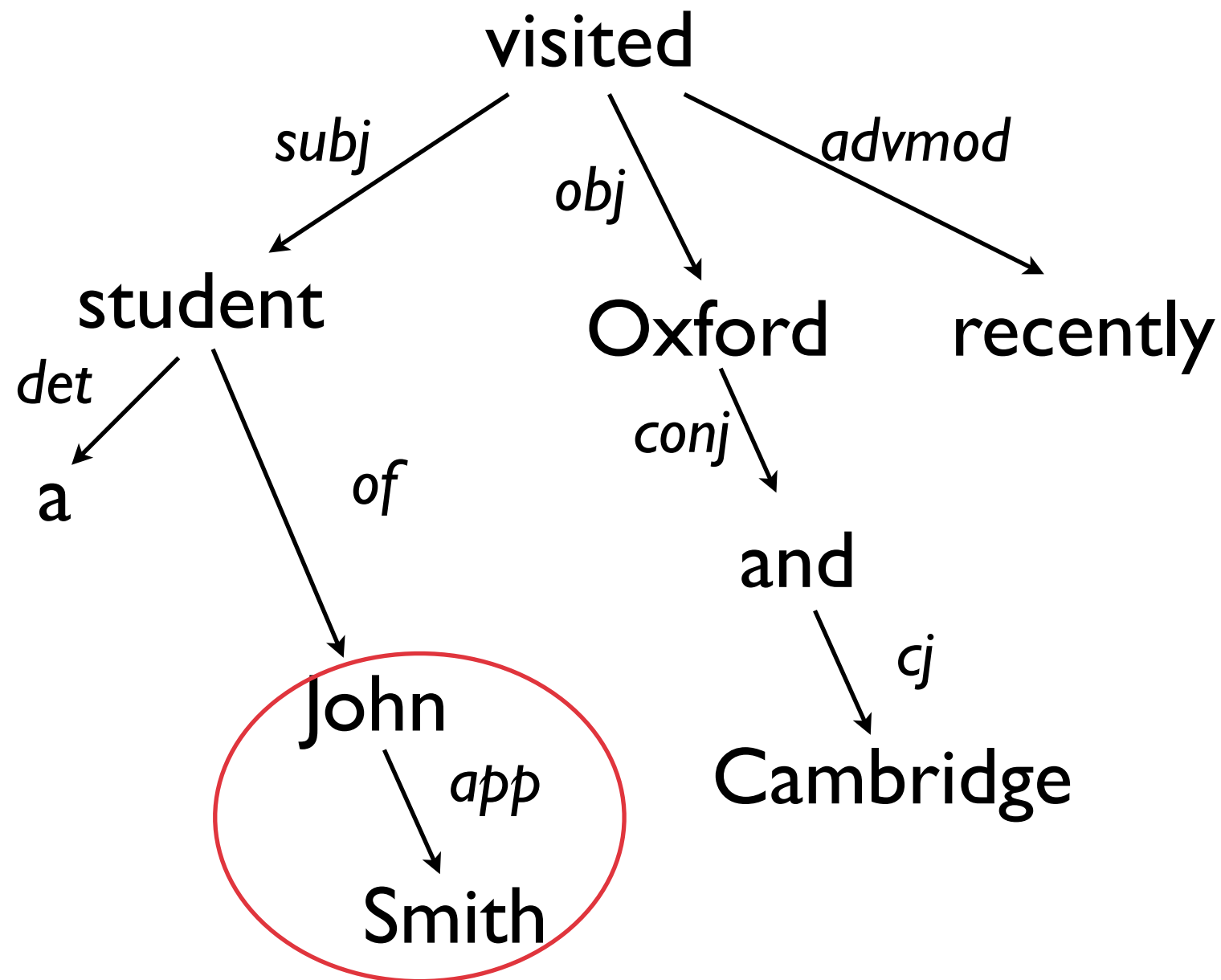
iii. Sentence fusion



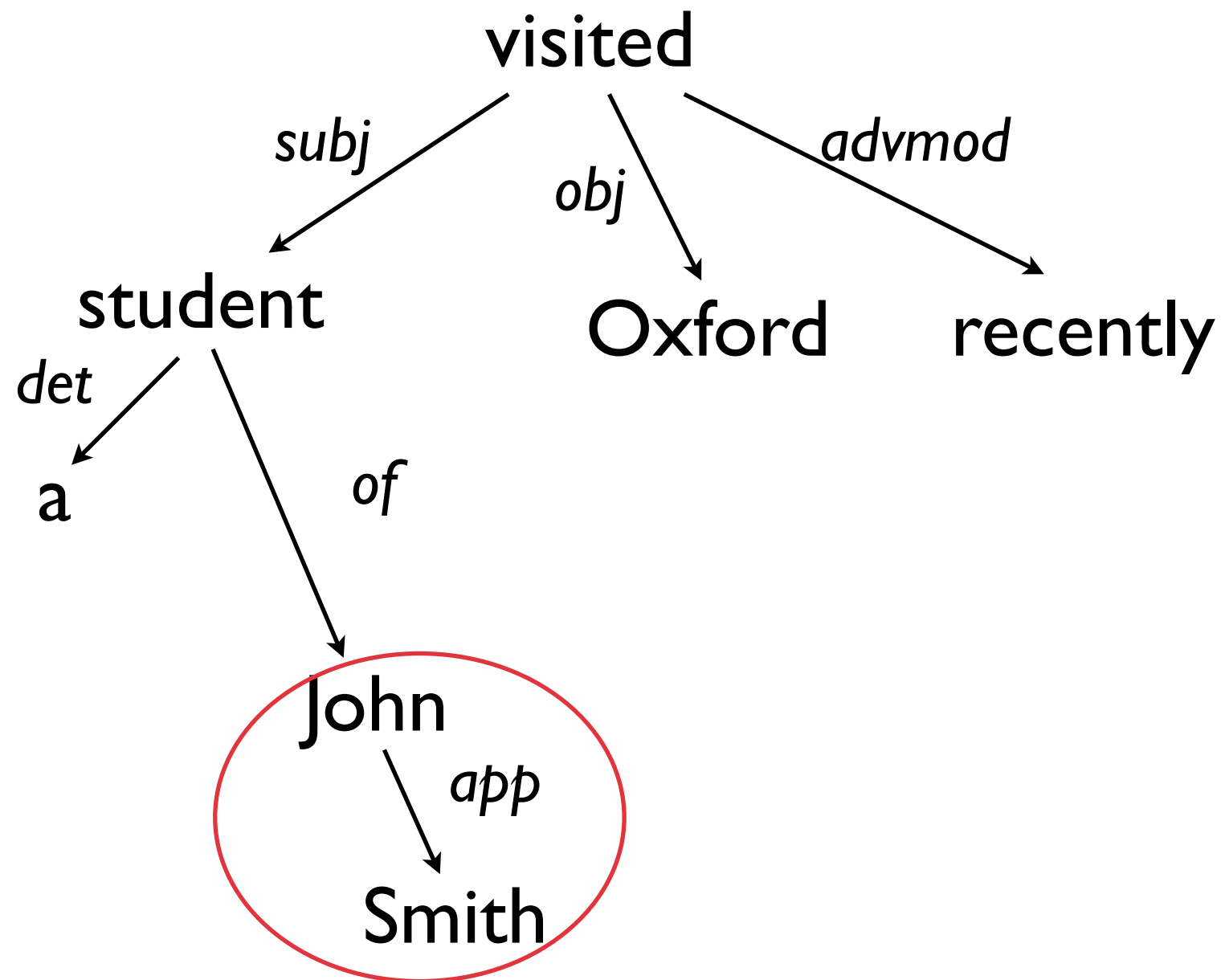
iii. Sentence fusion



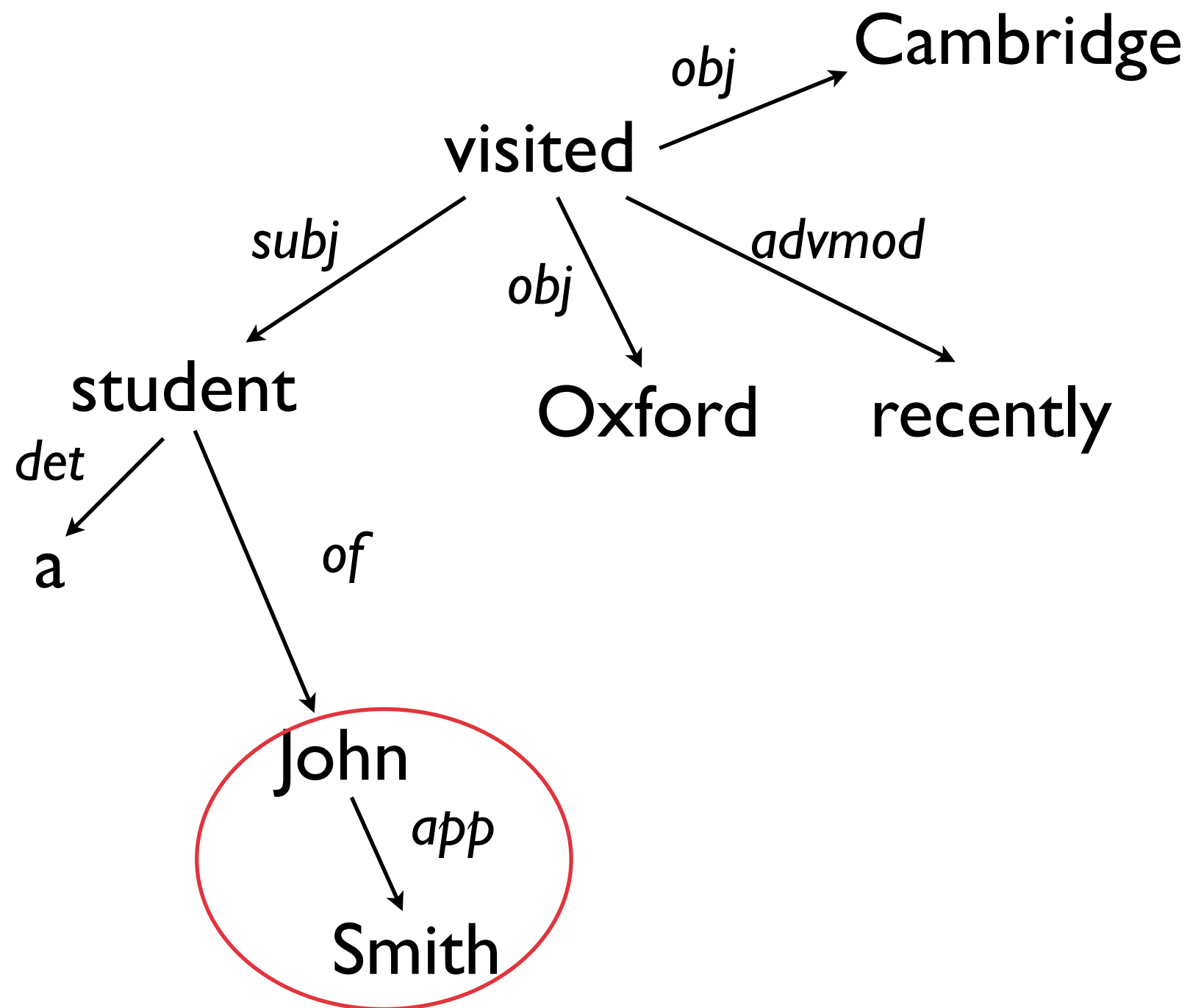
iii. Sentence fusion



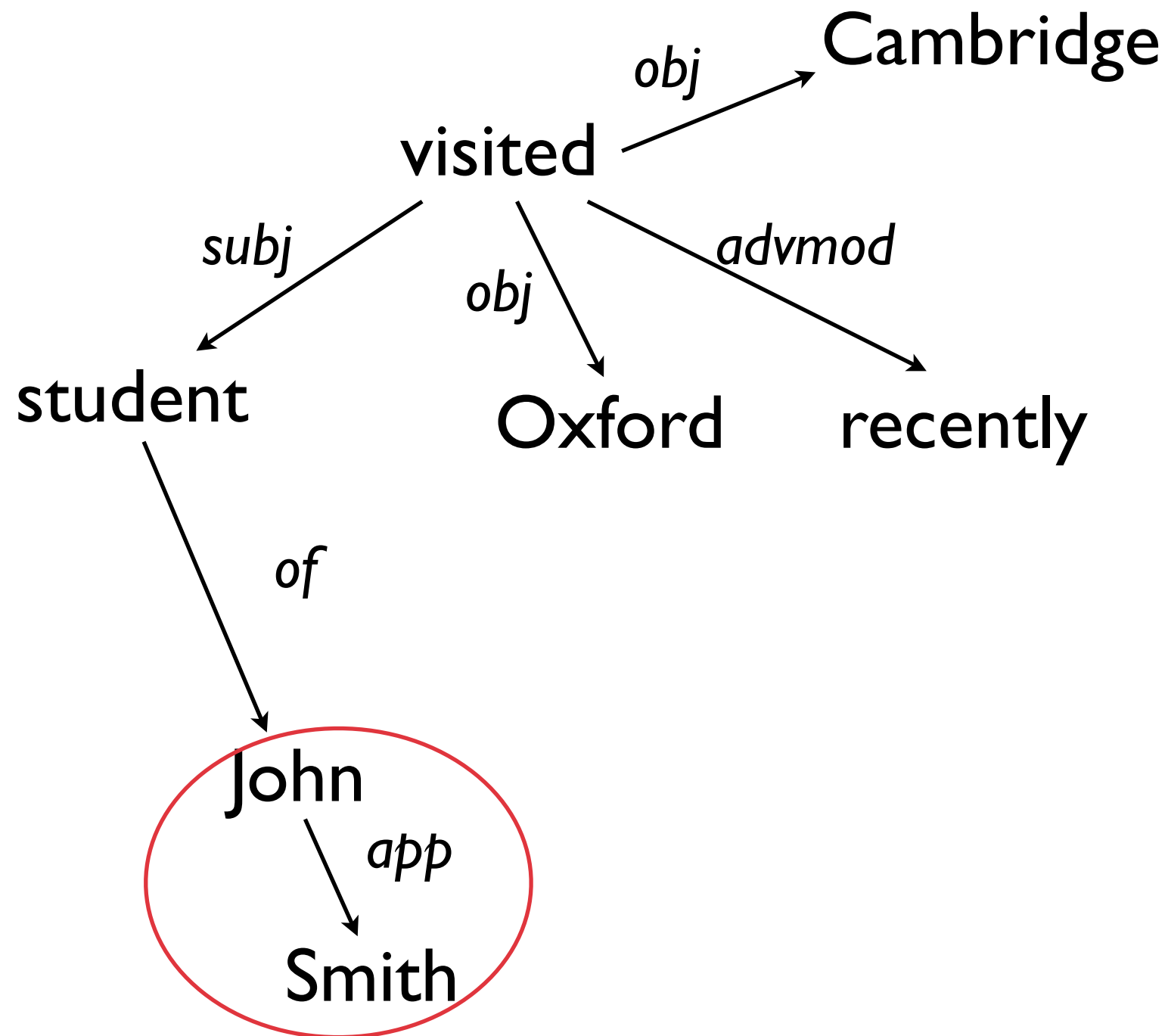
iii. Sentence fusion



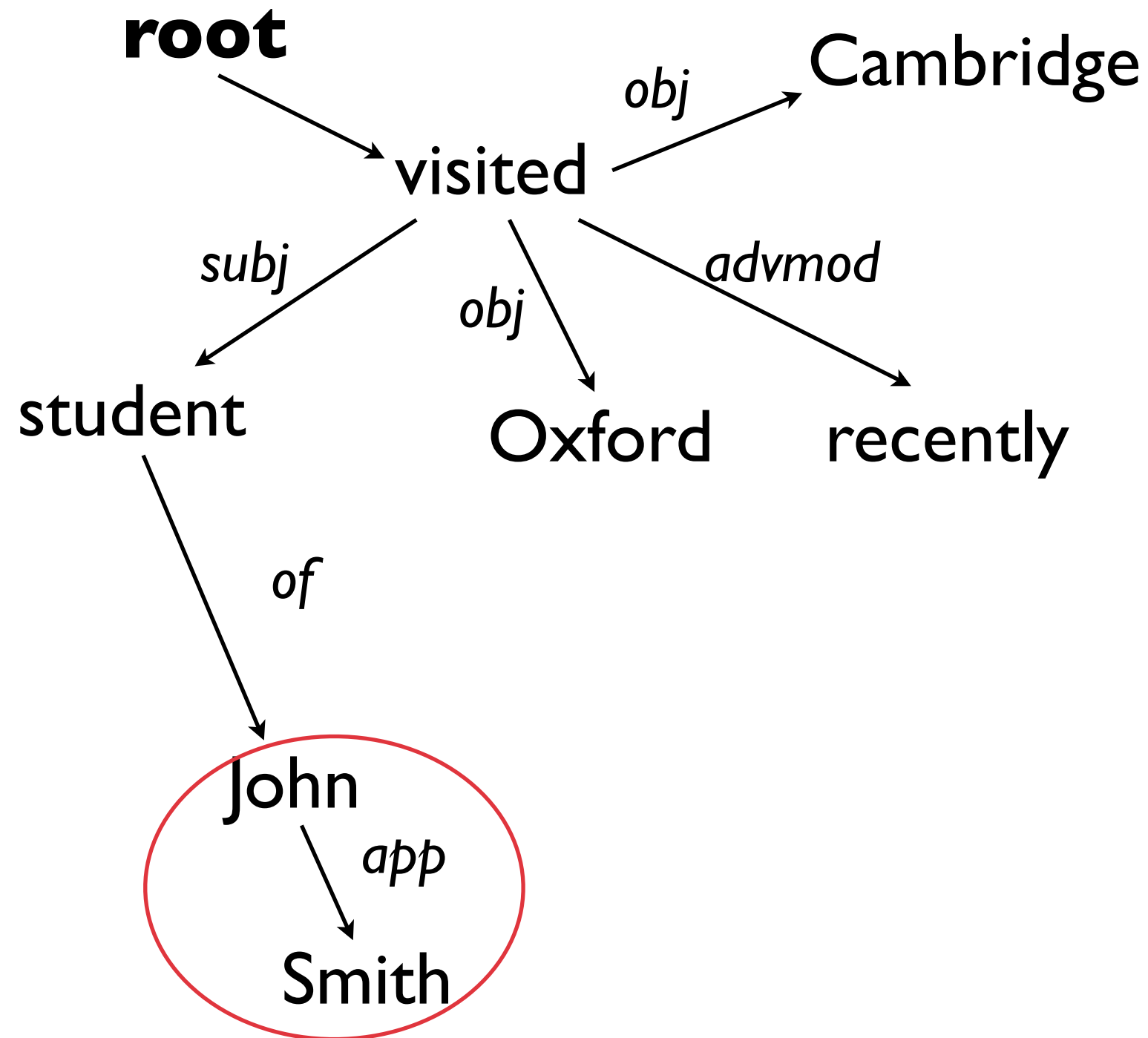
iii. Sentence fusion



iii. Sentence fusion



iii. Sentence fusion



iii. Sentence fusion

- Given such modified dependency representations of related sentences, join them in a single DAG by merging identical words, synonyms (e.g., WordNet) and entities (you can use Freebase, NE, coreference resolution).
- The resulting DAG covers all the input trees.
- Multiple dependency trees can be extracted from it, very few make sense.
- How can we find the best dependency tree?
- How can we find a valid / grammatical dependency tree?



iii. Sentence fusion

Several calls to tighten gun laws and monitor gun owners' accordance with storage requirement have been issued by politicians and other groups after 17-year-old Tim K., armed with a Beretta gun taken from his father's bedroom, killed 16 people in the small southwestern town of Winnenden, near Stuttgart.

Kretschmer shot many of his victims in the head with his father's legally registered Beretta.

Authorities say 17-year-old Tim Kretschmer used one of his father's weapons to gun down 15 people in a rampage that began at his former high school Wednesday.

Kretschmer gunned down students and teachers at his former high school before fleeing on foot and by car, killing three more people, and eventually shooting himself in the head, police said.



Sentence fusion



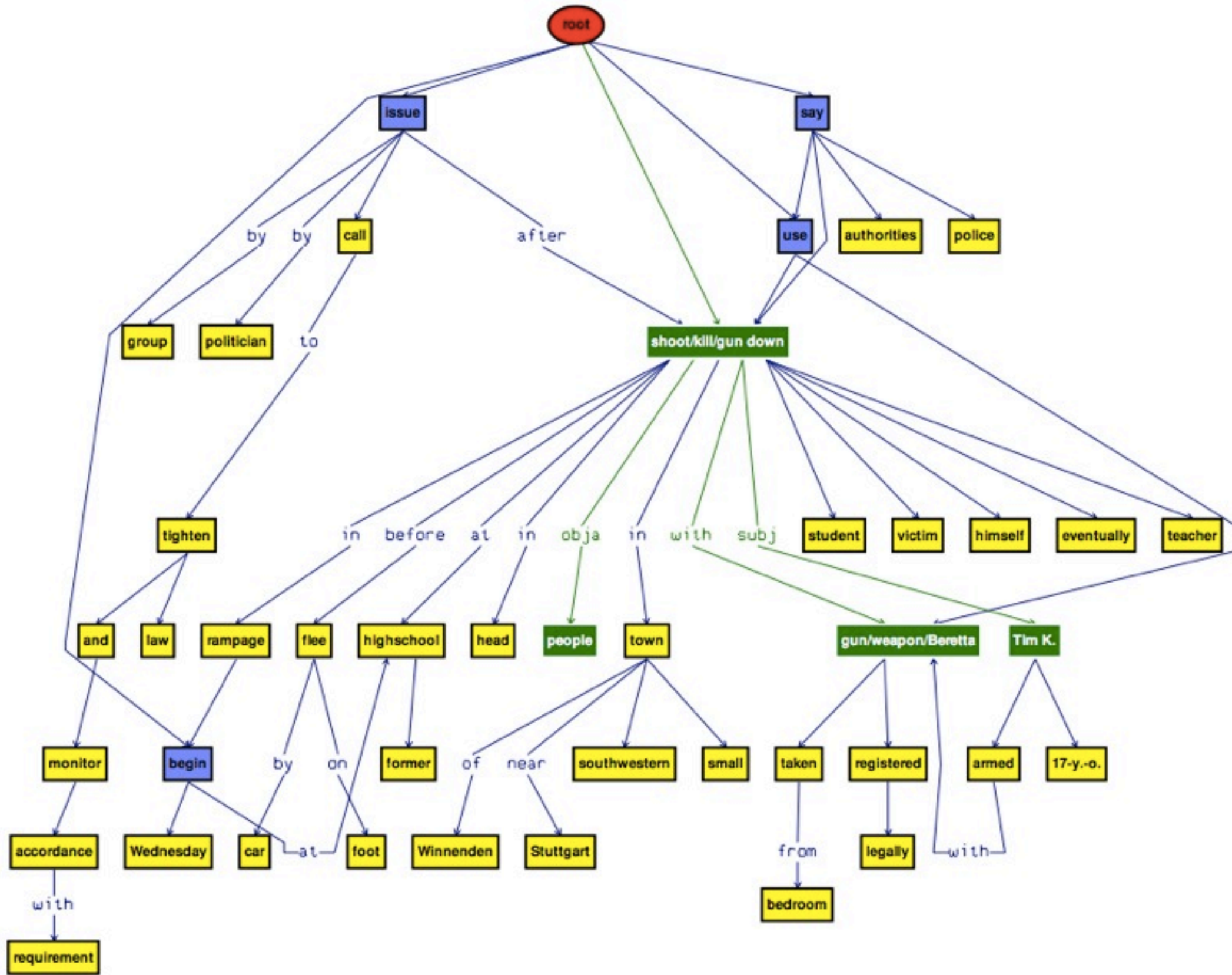
iii. Sentence fusion

Short “intersection” sentence:

Tim K. killed 16 people with his father’s gun.



iii. Sentence fusion

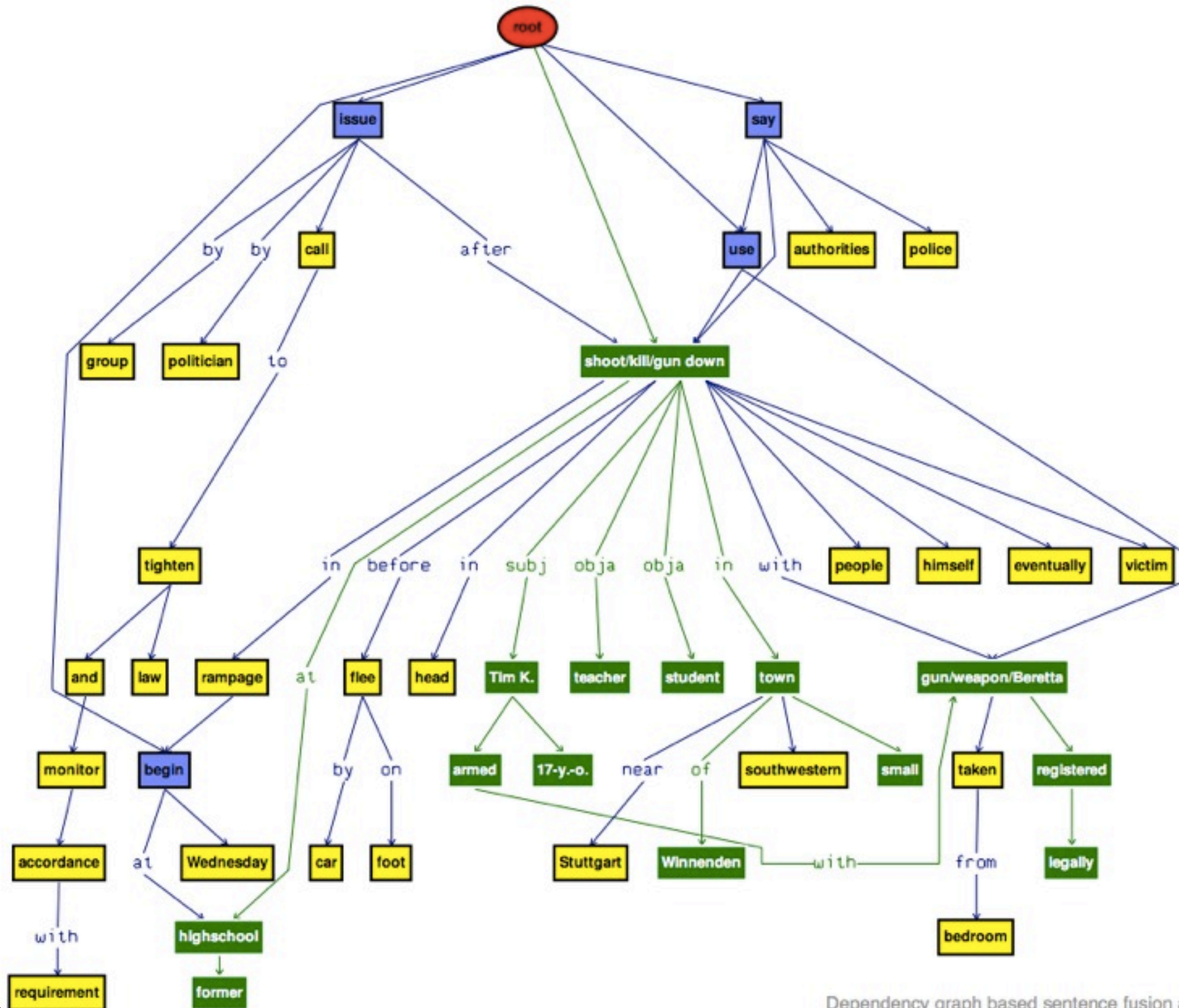


iii. Sentence fusion

More informative sentence:

17-years-old Tim K., armed with his father's legally registered gun, killed students and teachers at his former high-school in the small town of Winnenden.

iii. Sentence fusion



Dependency graph based sentence fusion analysis

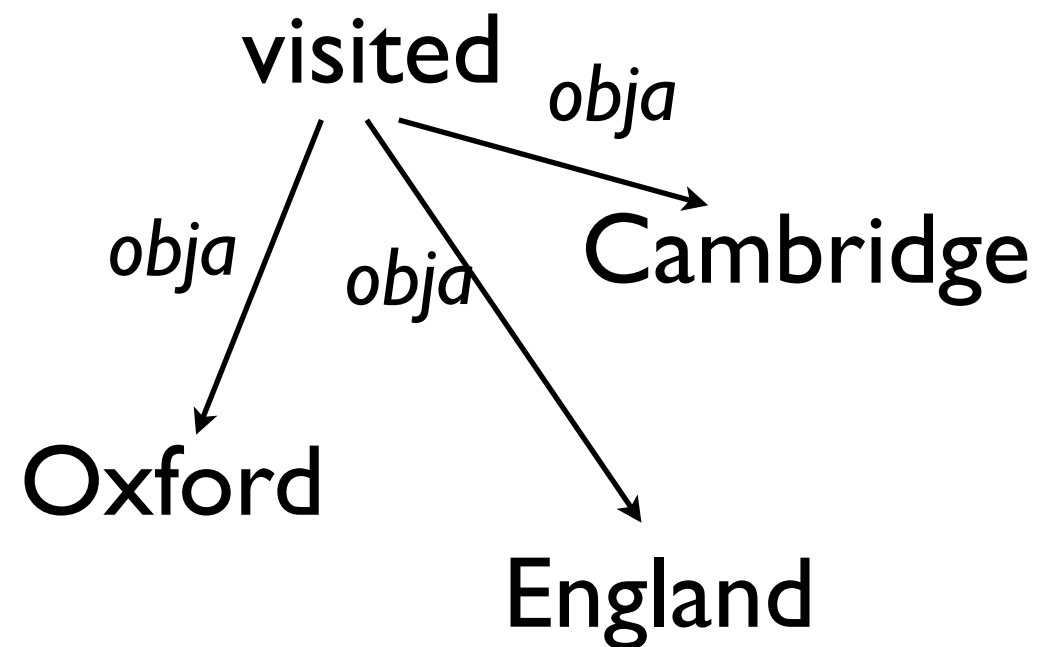
iii. Sentence fusion

- We can use ILP to obtain grammatical and informative trees:
 - for every edge, introduce a binary variable;
 - structural constraints to get a tree and not a random set of edges;
 - we can add syntactic, semantic, discourse constraints.
- But what are edge weights? Which edges are more important? $p(\text{label} \mid \text{lex-head})$ as a measure of syntactic importance, $MLE_{\text{estimated}}$.
 - no need to use lexicons or rules like in previous work;
 - all is needed is a parsed corpus.



iii. Sentence fusion

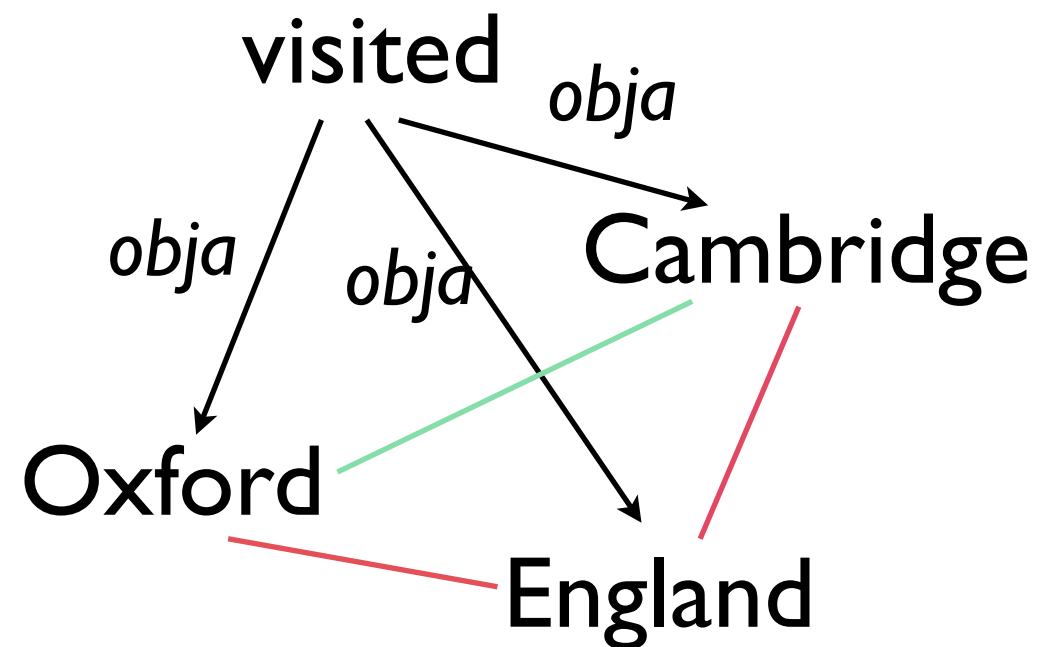
- Examples of semantic constraints:
 - do not retain take more than one edge from the same parent with the same label if the dependents are in ISA relation.



- do not retain two edges from the same head and with the same label if the lexical similarity between dependents is low: “*studies with pleasure and Niels Bohr*”, $\text{sim}(\text{pleasure}, \text{N.B}) = 0.01$.

iii. Sentence fusion

- Examples of semantic constraints:
 - do not retain take more than one edge from the same parent with the same label if the dependents are in ISA relation.



- do not retain two edges from the same head and with the same label if the lexical similarity between dependents is low: “*studies with pleasure and Niels Bohr*”, $\text{sim}(\text{pleasure}, \text{N.B}) = 0.01$.

iii. Sentence fusion

- What we have at this point is a dependency tree which still needs to be **linearized** - converted into a sentence, a string of words in a correct order.
 - we can overgenerate and rank again,
 - we can use a more efficient method ... [not presented here].
- A bonus: we can use the exact same method for sentence compression! [results comparable with state-of-the-art models on the mentioned datasets from C&L.]



iii. Sentence fusion

- Is the problem solved now? Not quite.

Q7: What problems / open questions do you see?



iii. Sentence fusion

- Is the problem solved now? Not quite.

Q7: What problems / open questions do you see?

- we can only generate words and dependencies seen in the input.
- we generate isolated sentences - would they fit together in a summary?
- how can we integrate world knowledge, e.g., add background information?



iii. Sentence fusion

Questions?



iv. Question generation

- A text-to-text generation task: given a sentence, make a question out of it.

iv. Question generation

- A text-to-text generation task: given a sentence, make a question out of it.

True toads are widespread and occur natively on every continent except Australia and Antarctica, inhabiting a variety of environments, from arid areas to rainforest.



iv. Question generation

- A text-to-text generation task: given a sentence, make a question out of it.

True toads are widespread and occur natively on every continent except Australia and Antarctica, inhabiting a variety of environments, from arid areas to rainforest.

How to find a sequence of operations
to get a good question?



iv. Question generation

- Three steps of QG (Heilman & Smith, 2010):
 - sentence extraction and preprocessing;
 - rule-based syntax-driven answer-to-question transformations:
 - mark phrases that can't be answers,
 - pick an answer phrase, generate a question phrase,
 - verb transformations
 - statistical ranking of sentence quality: learn a regression model from feature representation of the question-answer pair to a score (as if given by a human).

iv. Question generation

- Example from Heilman & Smith 2010:

Monrovia was named after James Monroe, who was president of the United States in 1922.

iv. Question generation

- Example from Heilman & Smith 2010:

Monrovia was named after James Monroe, who was president of the United States in 1922.

Monrovia was named after James Monroe.

iv. Question generation

- Example from Heilman & Smith 2010:

Monrovia was named after James Monroe, who was president of the United States in 1922.

Monrovia was named after James Monroe.

Was Monrovia named after James Monroe.



iv. Question generation

- Example from Heilman & Smith 2010:

Monrovia was named after James Monroe, who was president of the United States in 1922.

Monrovia was named after James Monroe.

Was Monrovia named after James Monroe.

Was Monrovia named after who.



iv. Question generation

- Example from Heilman & Smith 2010:

Monrovia was named after James Monroe, who was president of the United States in 1922.

Monrovia was named after James Monroe.

Was Monrovia named after James Monroe.

Was Monrovia named after who.

Who was Monrovia named after?



iv. Question generation

Questions?



Wrap-up



- Text-to-text generation - an open class of NLP tasks where the input and the output are text, e.g., paraphrase and question generation, sentence compression and fusion, text simplification.
- Shared representations, e.g., word lattices or dependency trees/graphs.
- Common formalisms, e.g., synchronous grammars, tree-edit models.
- Common approaches and techniques, e.g., noisy-channel, ILP, supervised ML from similar signals.



Pointers

- D2T generation:
 - GRE corpus: <http://www.csd.abdn.ac.uk/research/tuna/>
 - GRE Literature: <http://web.science.mq.edu.au/~jviethen/research.html>
 - GIVE: <http://www.give-challenge.org/research/>
- T2T generation:
 - <http://sites.google.com/site/t2tgen>



Pointers

- Paraphrasing:
 - Corpus: http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_corpus.html
 - Corpus/software: <http://www.cs.jhu.edu/~ccb/howto-extract-paraphrases.html>
 - Microsoft corpus: <http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/>
 - Collection of automatically extracted paraphrases: <http://www.cs.cornell.edu/Info/Projects/NLP/statpar.html>
 - Dutch parallel treebank: <http://www.inl.nl/en/corpora/daeso-corpus>



Pointers

- Sentence compression:
 - Broadcast/written: <http://jamesclarke.net/research/resources/>
 - Abstractive: <http://staffwww.dcs.shef.ac.uk/people/T.Cohn/t3/#Corpus>



Pointers

- Sentence fusion:
 - English news data: <http://www.cs.columbia.edu/~kapil/#turkfusion>
 - English review data: <http://kavita-ganesan.com/opinosis-opinion-dataset>
 - Dutch news: <http://daeso.uvt.nl/dutch-sentence-fusion-data/index.html>
 - German biographies: <http://www.h-its.org/english/research/nlp/download/cocobi.php>



Pointers



- Other resources:
 - PropBank: <http://verbs.colorado.edu/~mpalmer/projects/ace.html>
 - WordNet: <http://wordnet.princeton.edu>
 - FrameNet: <http://framenet.icsi.berkeley.edu>

