Word Alignment via Submodular Maximization over Matroids

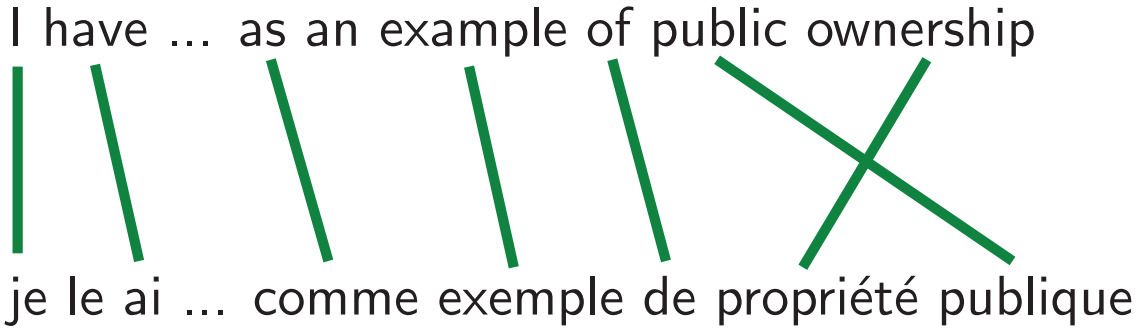
Hui Lin, Jeff Bilmes

University of Washington, Seattle
Dept. of Electrical Engineering

June 21, 2011

Matching-based Word Alignment

- Word alignment is a key component in most statistical machine translation systems
- Matching-based approaches (Melamed, 2000; Matusov et al., 2004; Taskar et al., 2005)
  - each word pair is associated with a score
  - alignment: highest scored matching



Our approach

We generalize

- linear (modular) objective $\rightarrow$ **submodular** objective
- matching constraint $\rightarrow$ **matroid** constraint

Benefit:

- expressive power (can represent complex interactions among alignment decisions).
- fertility of words are allowed to be larger than 1

Our approach

We generalize

- linear (modular) objective $\rightarrow$ **submodular** objective
- matching constraint $\rightarrow$ **matroid** constraint

Benefit:

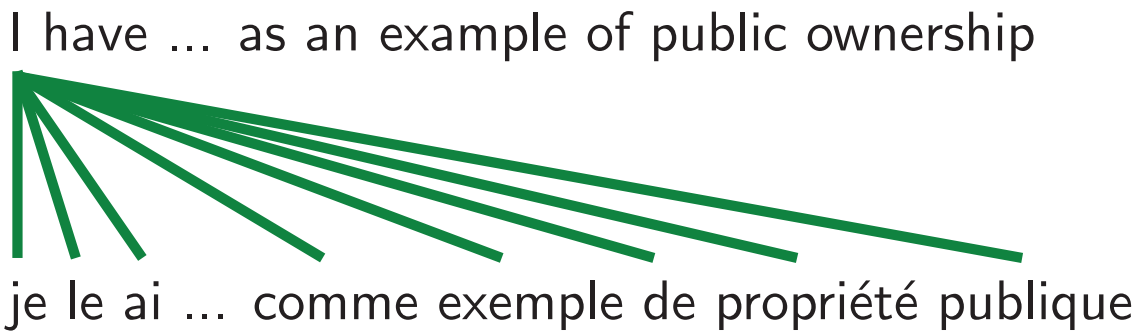
- expressive power (can represent complex interactions among alignment decisions).
- fertility of words are allowed to be larger than 1
- scalable and near-optimal algorithm is available, thanks to submodularity

Outline

Notation

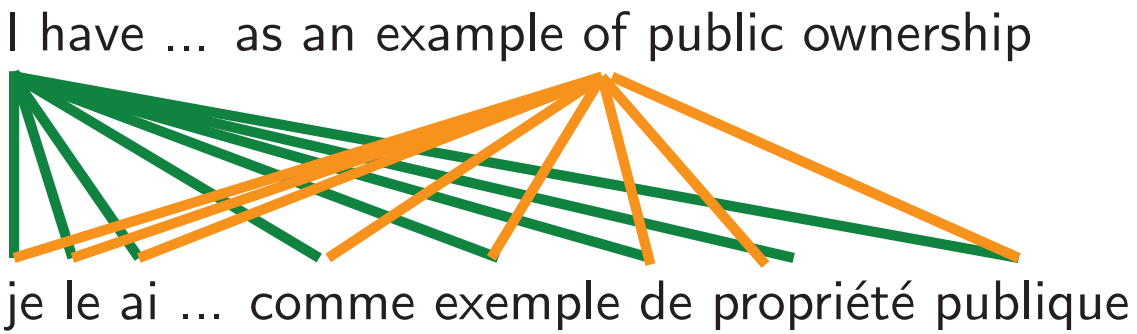
- A source language (English) string $e_1^I = e_1, \cdots, e_i, \cdots, e_I$
- A target language (French) string $f_1^J = f_1, \cdots, f_j, \cdots, f_J$
- $E \triangleq \{1, \cdots, I\}$: word positions in the English string
- $F \triangleq \{1, \cdots, J\}$: word positions in the French string
- Ground set: $V = E \times F = \{(i,j) : i \in E, j \in F\}$, set of edges in a bipartite graph.
- Any alignment A is a subset of V.
- $s_{i,j}$: score of aligning word i and j.

Partition of Word Alignments



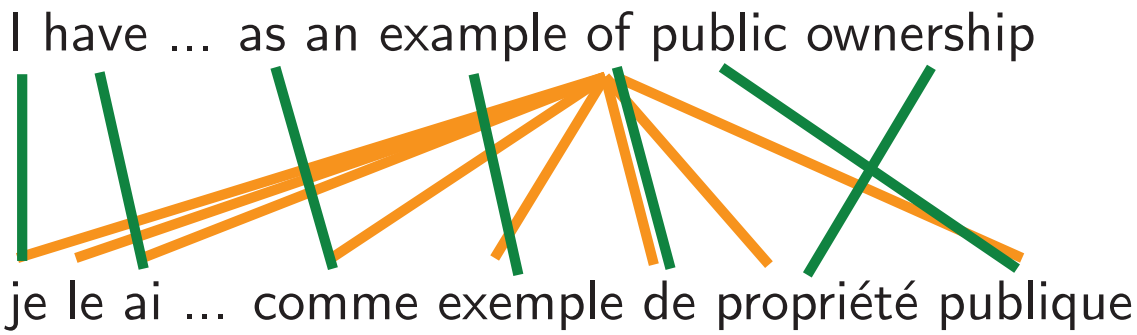
- for English word at position i, define $P_i^F = \{i\} \times F$
- P_i^F: the set of all possible edges in the ground set that connect to i

Partition of Word Alignments



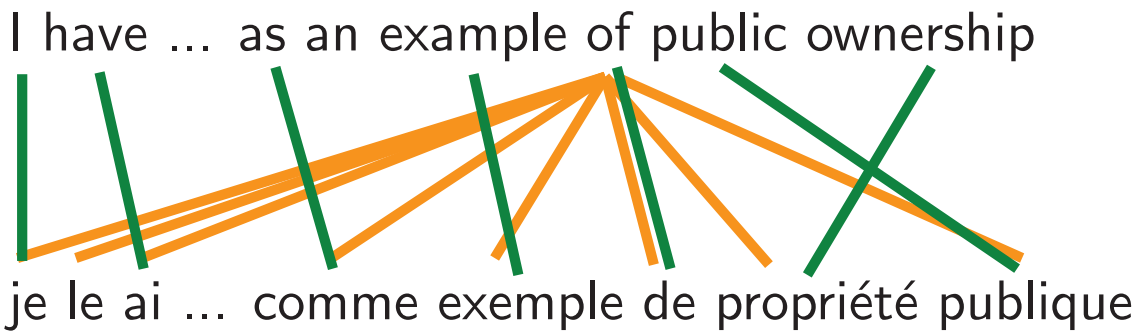
- for English word at position i, define $P_i^F = \{i\} \times F$
- P_i^F: the set of all possible edges in the ground set that connect to i
- for any $i \neq j$, $P_i^F \cap P_j^F = \emptyset$, and $\bigcup_{i \in E} P_i^F = V$
- therefore, $P_i^F, \forall i \in E$ forms a **Partition** of the ground set V.

Partition Matroid for Word Alignments



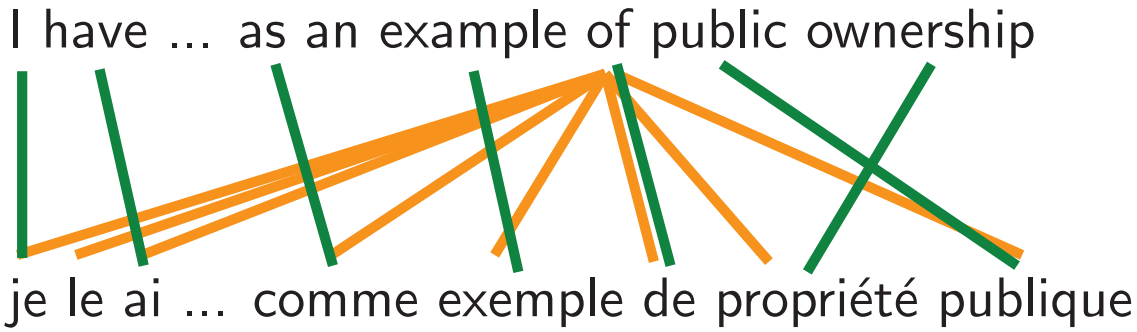
- $|A \cap P_i^F|$: how many edges in A are connected to i

Partition Matroid for Word Alignments



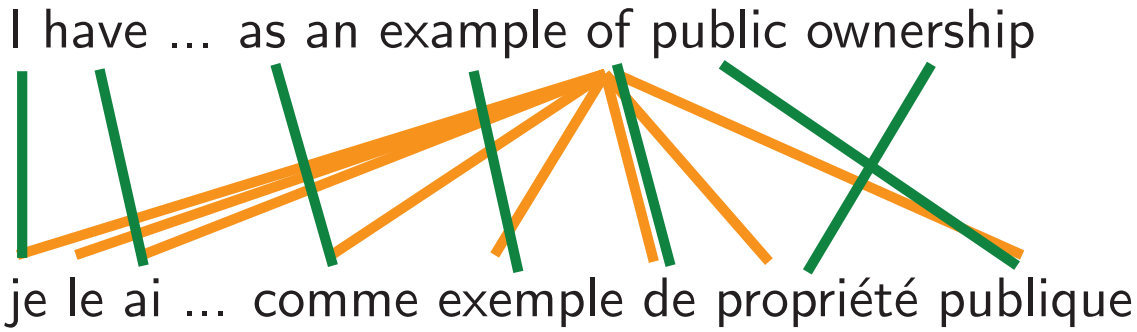
- $|A \cap P_i^F|$: how many edges in A are connected to i
- $|A \cap P_i^F| \leq k_i$: fertility of English word at i is at most k_i.

Partition Matroid for Word Alignments



- $|A \cap P_i^F|$: how many edges in A are connected to i
- $|A \cap P_i^F| \leq k_i$: fertility of English word at i is at most k_i.
- Constraint $\forall i \in E, |A \cap P_i^F| \leq k_i$ is a **partition matroid constraint**.

Partition Matroid for Word Alignments



- $|A \cap P_i^F|$: how many edges in A are connected to i
- $|A \cap P_i^F| \leq k_i$: fertility of English word at i is at most k_i.
- Constraint $\forall i \in E, |A \cap P_i^F| \leq k_i$ is a **partition matroid constraint**.
- A matroid is combinatorial structure that generalize the notion of **linear independence** in matrices, and it plays an important role in combinatorial optimization
- Example: sub-trees of a graph constitute a type of matroid.

Problem setup and algorithms

We use a set function $f : 2^V \rightarrow \mathbb{R}$ to model the quality (score) of an alignment A, then the word alignment problem can be cast as:

Problem

$$\max_{A \subseteq V} f(A), \text{ subject to:}$$

$$\forall i \in E, |A \cap P_i^F| \leq k_i,$$

$$\forall j \in F, |A \cap P_j^E| \leq k_j.$$

Problem setup and algorithms

We use a set function $f : 2^V \rightarrow \mathbb{R}$ to model the quality (score) of an alignment A, then the word alignment problem can be cast as:

Problem

$$\max_{A \subseteq V} f(A), \text{ subject to:}$$
$$\forall i \in E, |A \cap P_i^F| \leq k_i,$$
$$\forall j \in F, |A \cap P_j^E| \leq k_j.$$

- if $k_i = 1$ for all i, and $k_j = 1$ for all j, the constraints reduce to the standard matching constraints.
- if $f(S)$ is modular (linear) as well, the above problem is reduced to the problem solved in standard matching-based approaches.

Problem setup and algorithms

We use a set function $f : 2^V \rightarrow \mathbb{R}$ to model the quality (score) of an alignment A, then the word alignment problem can be cast as:

Problem

$$\max_{A \subseteq V} f(A), \text{ subject to:}$$
$$\forall i \in E, |A \cap P_i^F| \leq k_i,$$
$$\forall j \in F, |A \cap P_j^E| \leq k_j.$$

- if $k_i = 1$ for all i, and $k_j = 1$ for all j, the constraints reduce to the standard matching constraints.
- if $f(S)$ is modular (linear) as well, the above problem is reduced to the problem solved in standard matching-based approaches.

Problem setup and algorithms

We use a set function $f : 2^V \rightarrow \mathbb{R}$ to model the quality (score) of an alignment A, then the word alignment problem can be cast as:

Problem

$$\max_{A \subseteq V} f(A), \text{ subject to:}$$
$$\forall i \in E, |A \cap P_i^F| \leq k_i,$$
$$\forall j \in F, |A \cap P_j^E| \leq k_j.$$

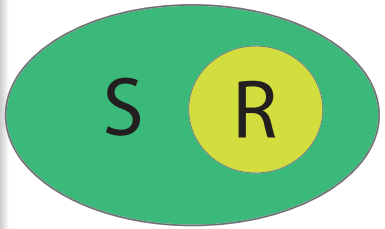
- Alternatively, If f is monotone submodular, a greedy algorithm can solve the above problem near-optimally even if k_i, k_j are not restricted to be 1 (see the paper for details).

Submodular Set Function

Definition of Submodular Functions

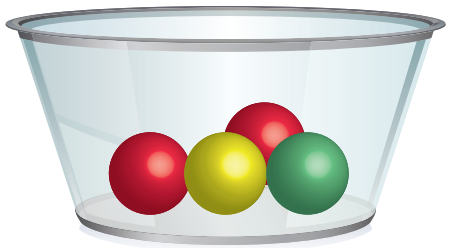
For any $R \subseteq S \subseteq V$ and $k \in V, k \notin S$, $f(\cdot)$ is **submodular** if

$$f(S + k) - f(S) \leq f(R + k) - f(R)$$

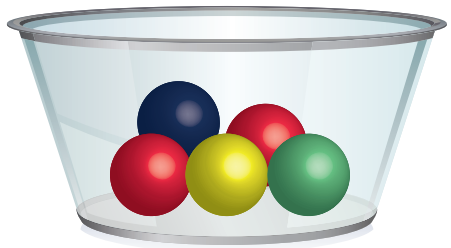


This is known as the principle of **diminishing returns**

Example: Number of Colors of Balls in Urns



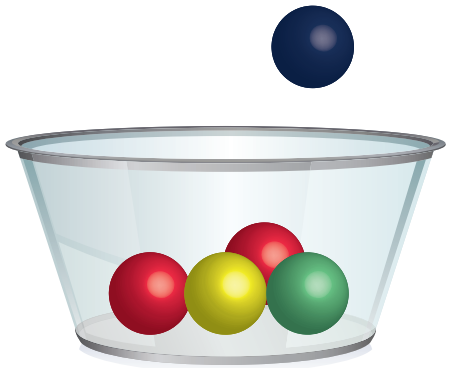
$f(R) = f($  $) = 3$



$f(S) = f($  $) = 4$

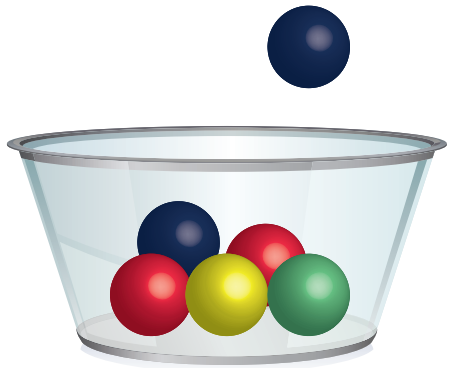
- Given a set A of colored balls
- $f(A)$: the number of distinct colors contained in the urn
- Incremental value of object **diminishes** in a **larger** context (diminishing returns).

Example: Number of Colors of Balls in Urns



$f(R) = f(\quad) = 3$

$f(R + k) = f(\quad + \quad) = 4$

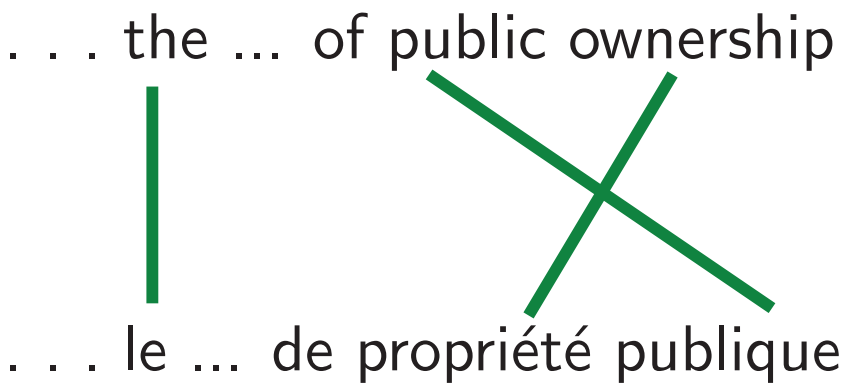


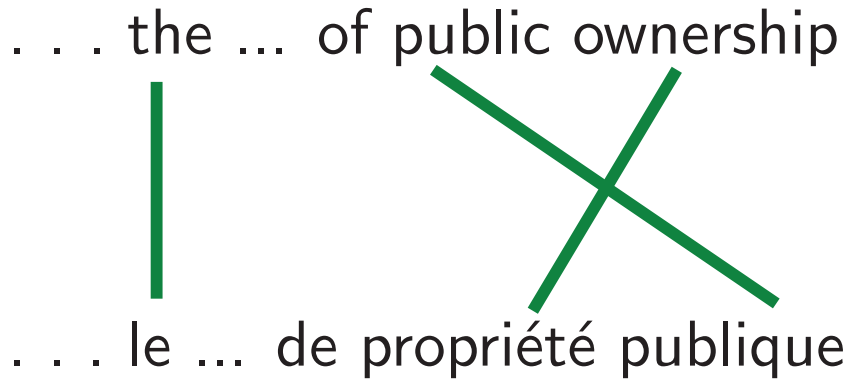
$f(S) = f(\quad) = 4$

$f(S + k) = f(\quad + \quad) = 4$

- Given a set A of colored balls
- $f(A)$: the number of distinct colors contained in the urn
- Incremental value of object **diminishes** in a **larger** context (diminishing returns).

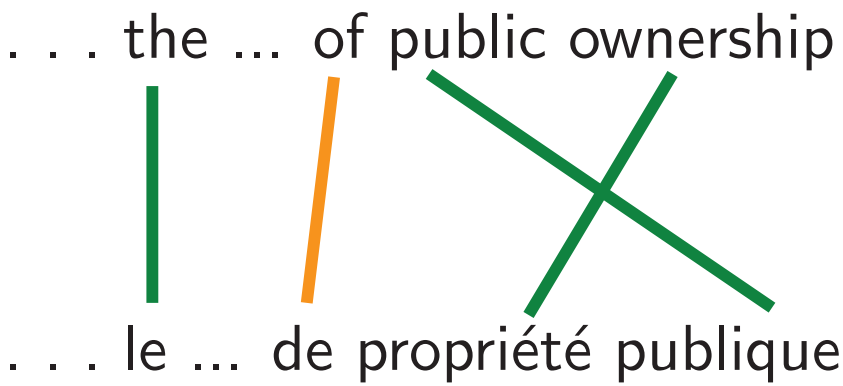
Submodular Fertility

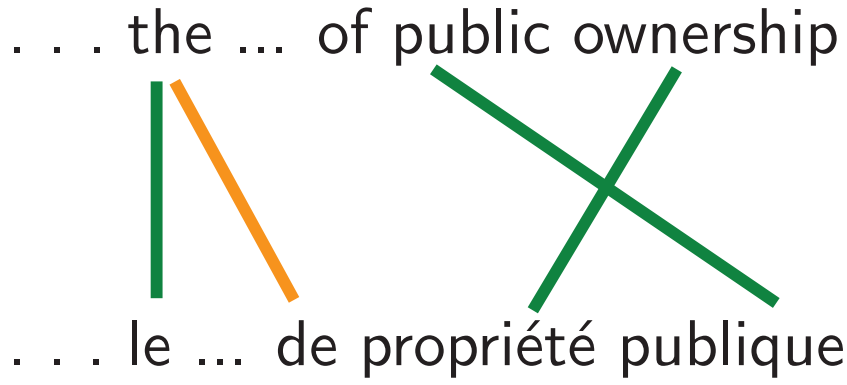




- consider align "de" to "the" or to "of"

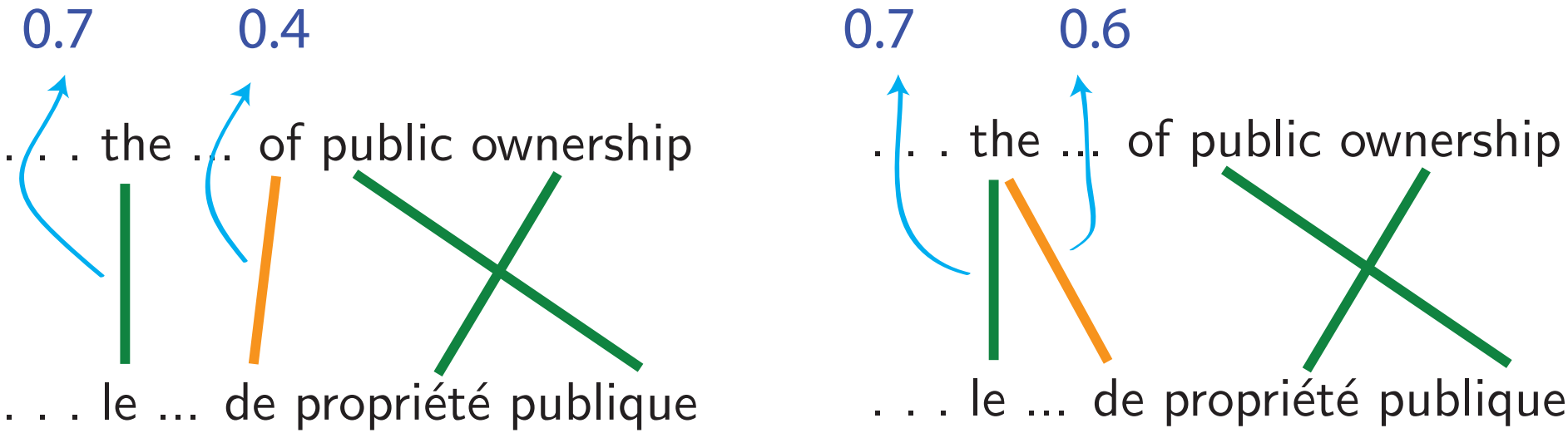
Submodular Fertility





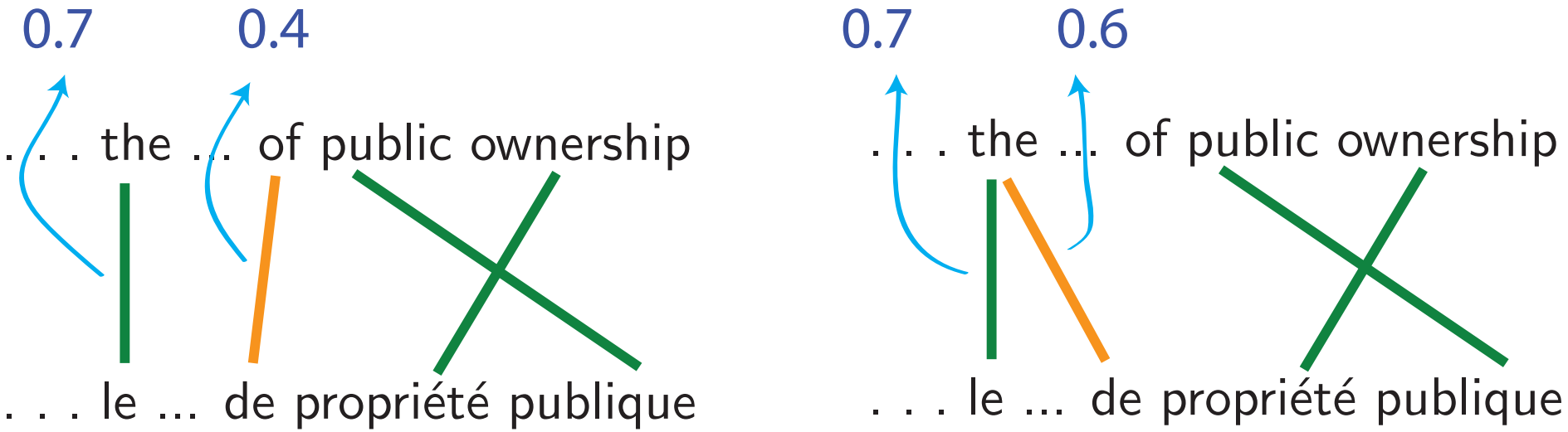
- consider align "de" to "the" or to "of"
- the correct alignment should be the one on the left

Submodular Fertility



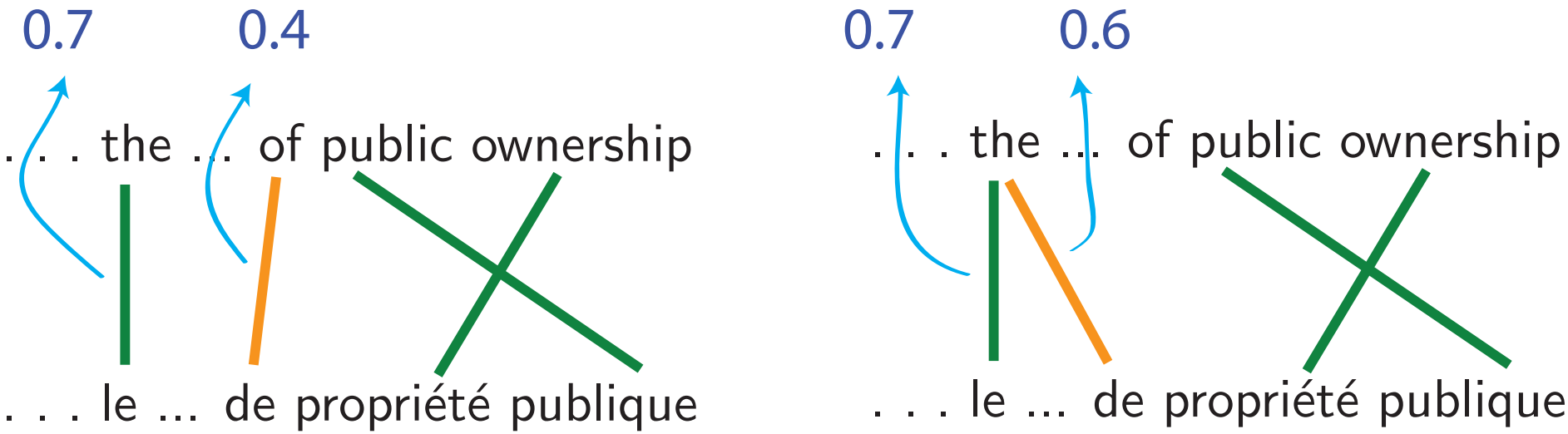
- consider align "de" to "the" or to "of"
- the correct alignment should be the one on the left
- modular (linear) objective: $0.7 + 0.4 < 0.7 + 0.6$.
- competitive linking (Melamed, 2000): one-to-one mapping assumption

Submodular Fertility



- consider align "de" to "the" or to "of"
- the correct alignment should be the one on the left
- modular (linear) objective: $0.7 + 0.4 < 0.7 + 0.6$.
- competitive linking (Melamed, 2000): one-to-one mapping assumption
- **diminishing** the marginal benefits of aligning a word to more than one words in the other string

Submodular Fertility



- modular (linear) objective: $0.7 + 0.4 < 0.7 + 0.6$.
- submodular objective: $0.7 + 0.4 > \sqrt{0.7 + 0.6}$.
- thus, a submodular objective gets the better alignment in this case

Results

- Data: English-French Hansards data from the 2003 NAACL shared task

Table: AER results

ID	Objective function	Constraint	AER(%)
1	modular	$\text{Fert}_F(A) \leq 1, \text{Fert}_E(A) \leq 1$	21.0
2		$\text{Fert}_F(A) \leq 1$	23.1
3		$\text{Fert}_F(A) \leq k_j$	22.1
4	submodular	$\text{Fert}_F(A) \leq 1$	19.8
5		$\text{Fert}_F(A) \leq k_j$	**18.6**
Generative model (IBM 2, E→F)			21.0
Maximum weighted bipartite matching			20.9
Matching with negative penalty on fertility (ILP)			19.3

Results

- Data: English-French Hansards data from the 2003 NAACL shared task

Table: AER results

ID	Objective function	Constraint	AER(%)
1	modular	$\text{Fert}_F(A) \leq 1, \text{Fert}_E(A) \leq 1$	21.0
2		$\text{Fert}_F(A) \leq 1$	23.1
3		$\text{Fert}_F(A) \leq k_j$	22.1
4	submodular	$\text{Fert}_F(A) \leq 1$	19.8
5		$\text{Fert}_F(A) \leq k_j$	**18.6**
Generative model (IBM 2, E→F)			21.0
Maximum weighted bipartite matching			20.9
Matching with negative penalty on fertility (ILP)			19.3

- our approach is at least 50 times faster than the ILP-based approach

Summary

- a novel framework where word alignment is framed as submodular maximization subject to matroid constraints.
- extension of previous matching-based frameworks in two respects:
  1. modular objective $\rightarrow$ submodular objective
  2. matching constraint $\rightarrow$ matroid constraint
- such generalizations do not incur a prohibitive computational cost since submodular maximization over matroids can be efficiently approximated with high quality performance guarantees.
- the full potential of our approach has yet to be explored: we plan to investigate richer submodular functions as well (see our paper/presentation "A Class of Submodular Functions for Document Summarization" from yesterday).
- our approach might lead to novel approaches for machine translation as well.