Evaluation Metrics for Automatic Temporal Annotation of Texts

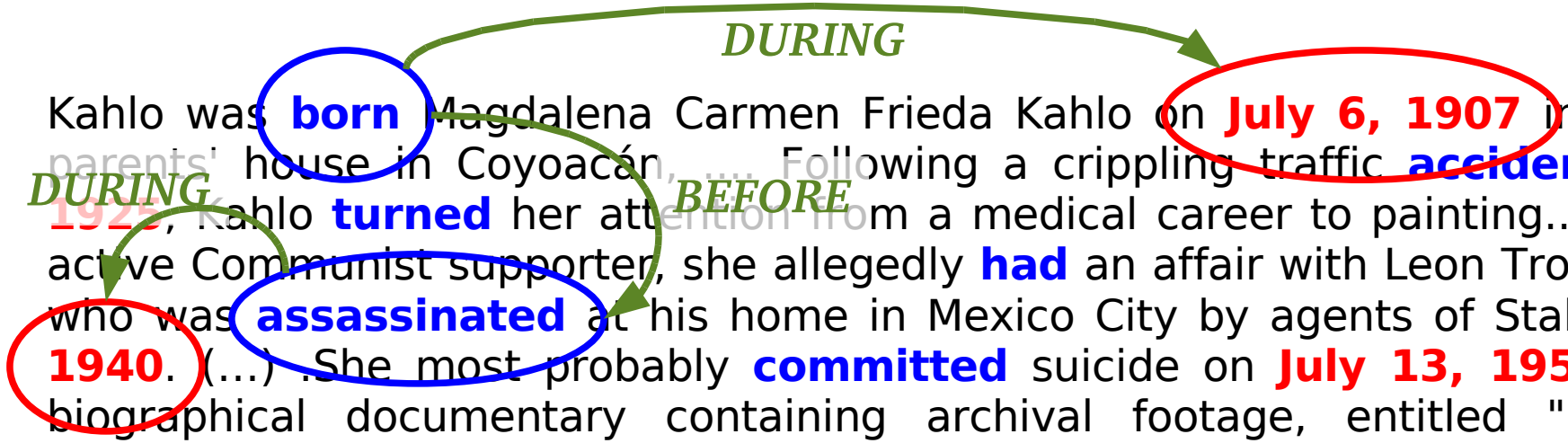
Xavier Tannier
LIMSI-CNRS
Orsay
France

Philippe Muller
IRIT
Toulouse
France

LREC 2008, May 29

What is automatic temporal annotation?

- Extracting events and temporal expressions from texts
- Time-stamping and temporal ordering



Kahlo was **born** Magdalena Carmen Frieda Kahlo on **July 6, 1907** in her parents' house in Coyoacán, ... Following a crippling traffic **accident** in **1925**, Kahlo **turned** her attention from a medical career to painting.... An active Communist supporter, she allegedly **had** an affair with Leon Trotsky, who was **assassinated** at his home in Mexico City by agents of Stalin in **1940**. (...) She most probably **committed** suicide on **July 13, 1954**. A biographical documentary containing archival footage, entitled "Frida Kahlo", was **released** in **1982**, in Germany. In **1984** director Paul Le Duc **released** the film "Frida, naturaleza viva", which stars Ofelia Medina as Frida Kahlo. In **2002**, **released** a motion picture titled "Frida", starring Salma Hayek in the title role.

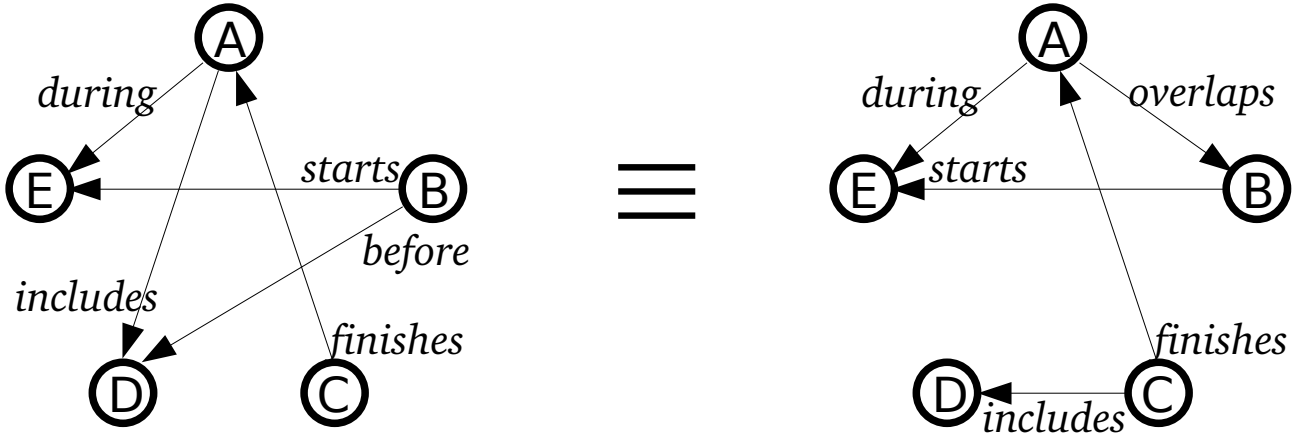
What do we evaluate?

- Temporal relations between:
  - Pairs of events
  - Events and calendar expressions

- TempEval 2007 at SemEval

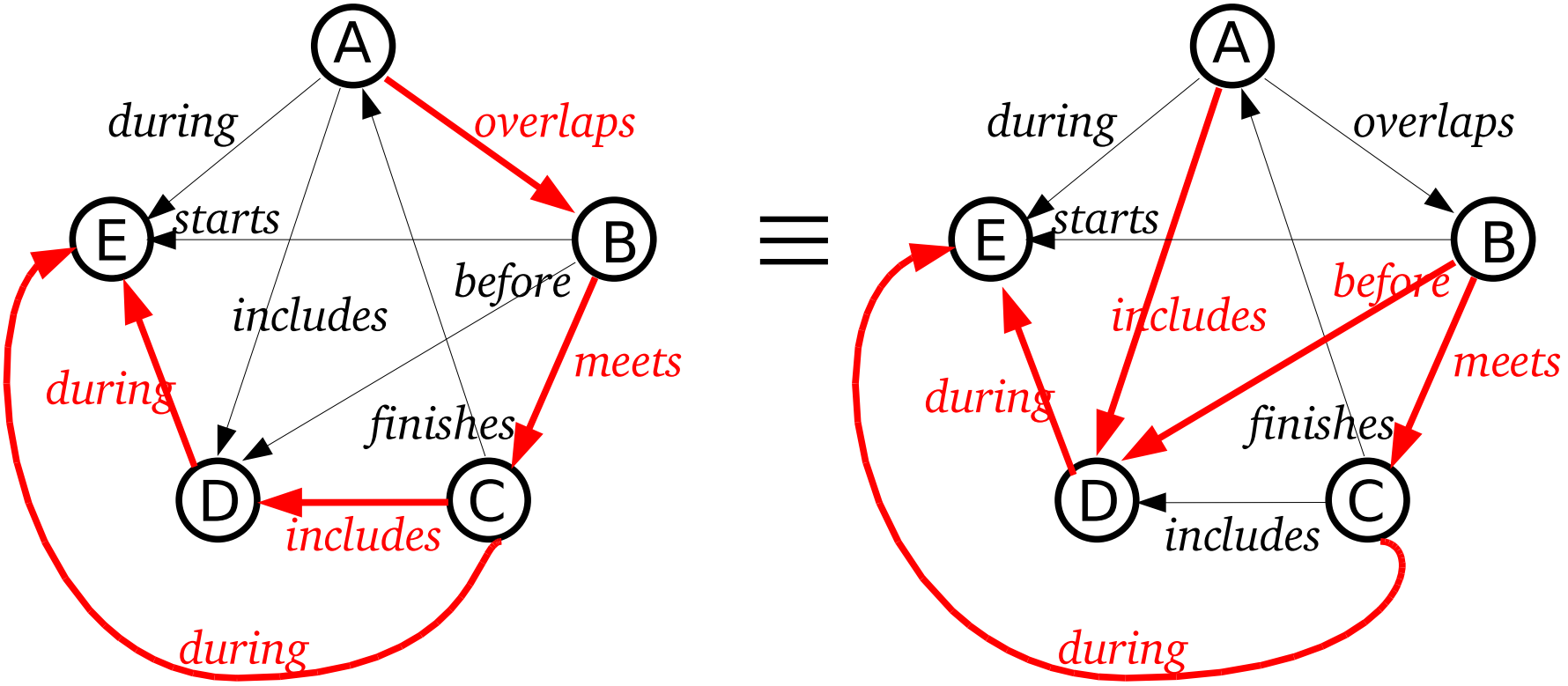
Why is the evaluation difficult?

- The set of relations varies among systems
- Vague (disjunctive) annotation handling is necessary
  - "**a is *before* b**" *vs.* "**a is *before* b OR a *overlaps* b**"
  - Precision and recall should be weighted through this particularity [Muller and Tannier, 2004]
- Equivalent relations can be expressed in different ways:

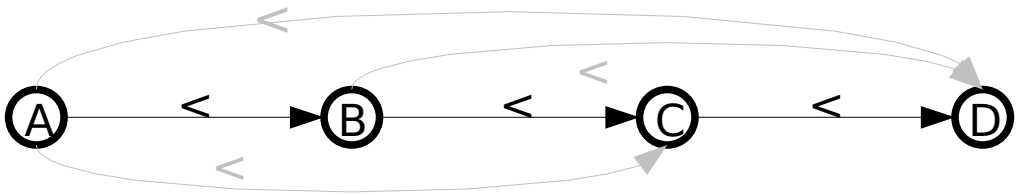
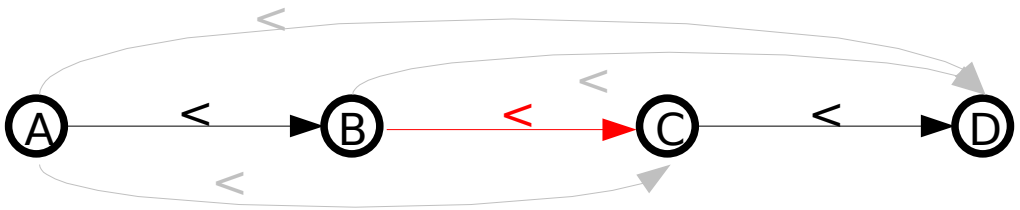
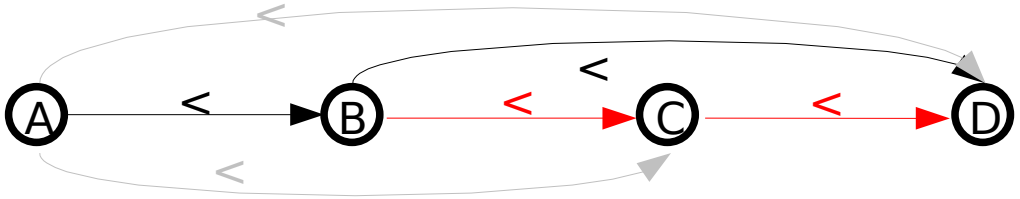
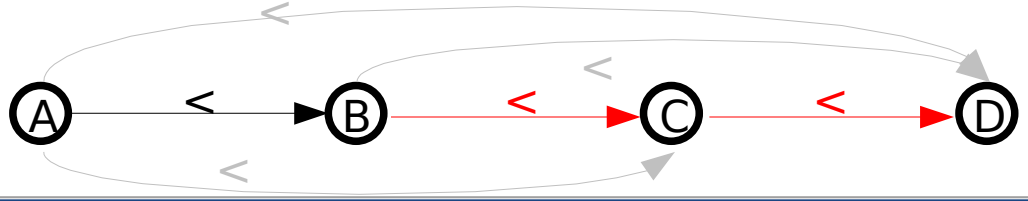


(example from [Rodríguez et al., 2004])

Temporal closure



Relative importance of relations

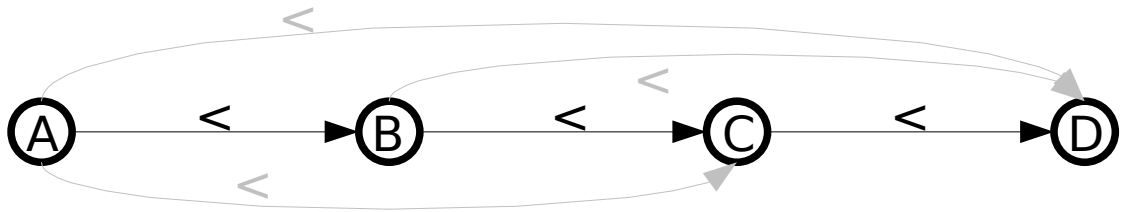
	Current metrics' recall	*Expected score*
	1	1
	$\frac{2}{6}$	$\approx \frac{2}{3}$
	$\frac{2}{6}$	$> \frac{1}{3}$
	$\frac{1}{6}$	$\approx \frac{1}{3}$

Problems

- That's unfair! ...
- ... Feeling that we formalize as:
  - For n nodes, up to $\frac{n \times n-1}{2}$ edges can hold in the graph
  - With usual metrics on temporal closure, the importance given to a text in a corpus is not a linear, but square function of its "size" (in terms of its number of events)...
  - ... It should be linear: in the evaluation of a whole corpus, the impact of the measure on a text should be in proportion to its size

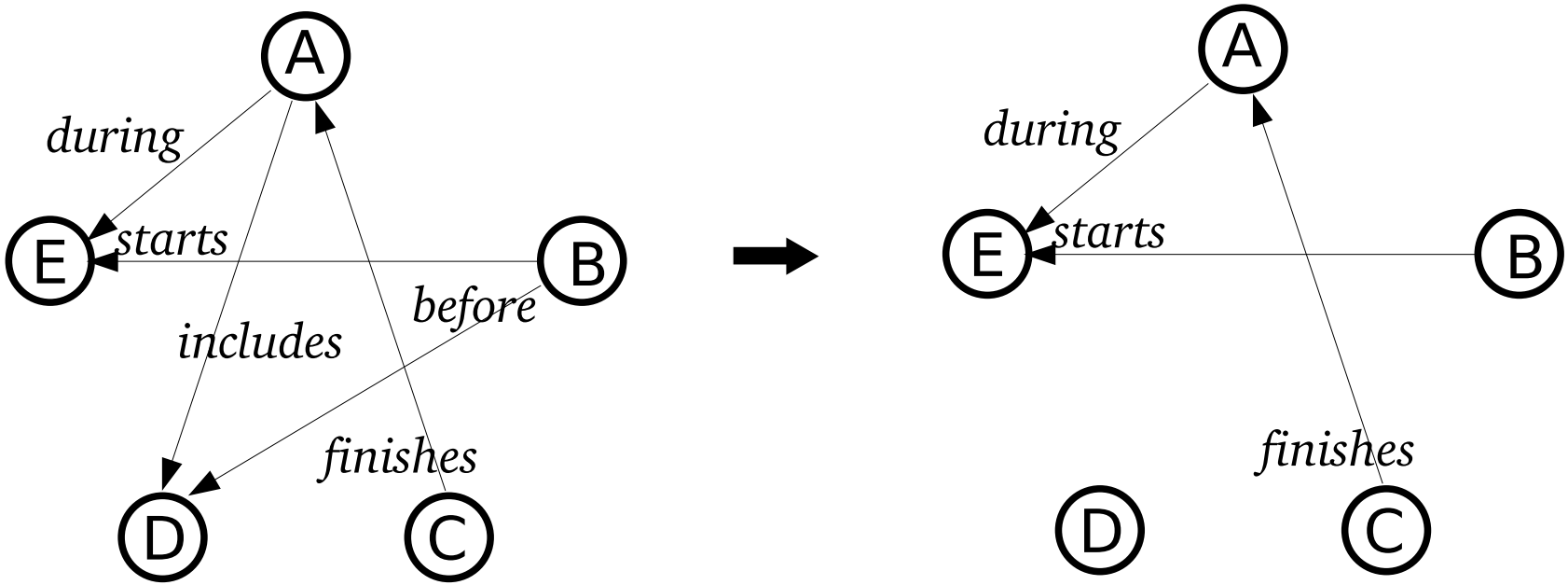
Minimal and kernel graphs

- *Minimal: A graph from which no relation can be removed without losing any temporal information after temporal closure*



- But in the general case, such a graph:
  - is not unique [Rodríguez et al., 2004]
  - is heavy to compute
- As an expedient, we use "kernel" relations:
  - Relations that lead to a loss of information when removed
  - Relations that appear in ALL minimal graphs
  - Unique and easy to compute, but not as much information

Example


A
during
E
starts
B
before
includes
finishes
D
C
A
during
E
starts
B
finishes
D
C

A demo metric

- We close both key (K) and candidate (G) graphs
- We look how many core relations from K have been found in G. This is kernel recall.
- We check how many core relations from G are also in K. This is kernel precision.

- Only an approximation of what should be assessed.
- But a good idea of how important the relations are in a same graph.
- We expect these measures to grow more linearly with the number of events in a text.

What a good metric should do?

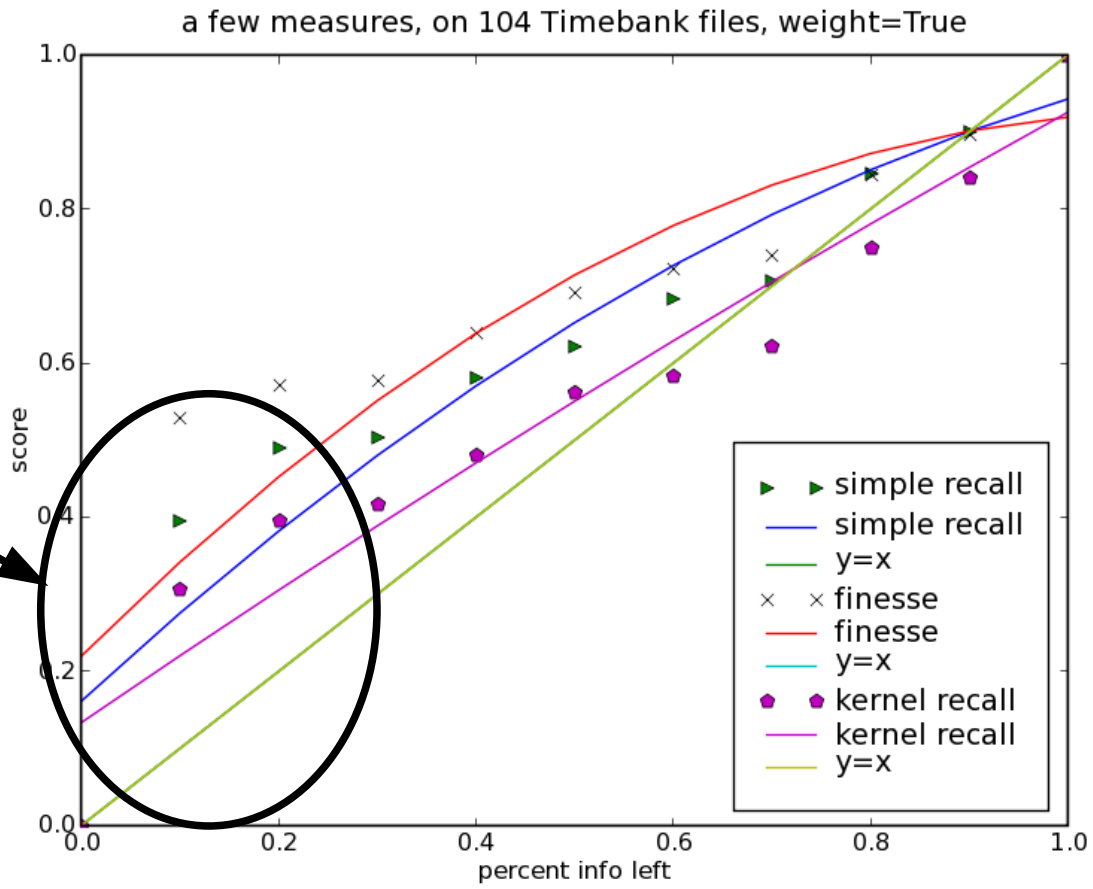
- Provide intuitive scores
- Decrease when information decreases (in a monotonic and regular way - and linearly with the level of information provided)
- Information can "decrease" when :
  - relations found are vaguer
  - less relations are found (decreasing recall)
  - more relations are inaccurate (decreasing precision)

Experiments

- Comparison of metrics:
  - Recall / Precision
  - Finesse / Cohesion [Muller and Tannier, 2004]
  - Kernel recall / kernel precision
- Study:
  - Removing temporal information one piece at a time (for recall)
  - Introducing noise (changing relation to another) (for precision)
- Corpus: TimeBank newswire corpus
- Expectation:
  - A good measure should decrease linearly with removal/change of information

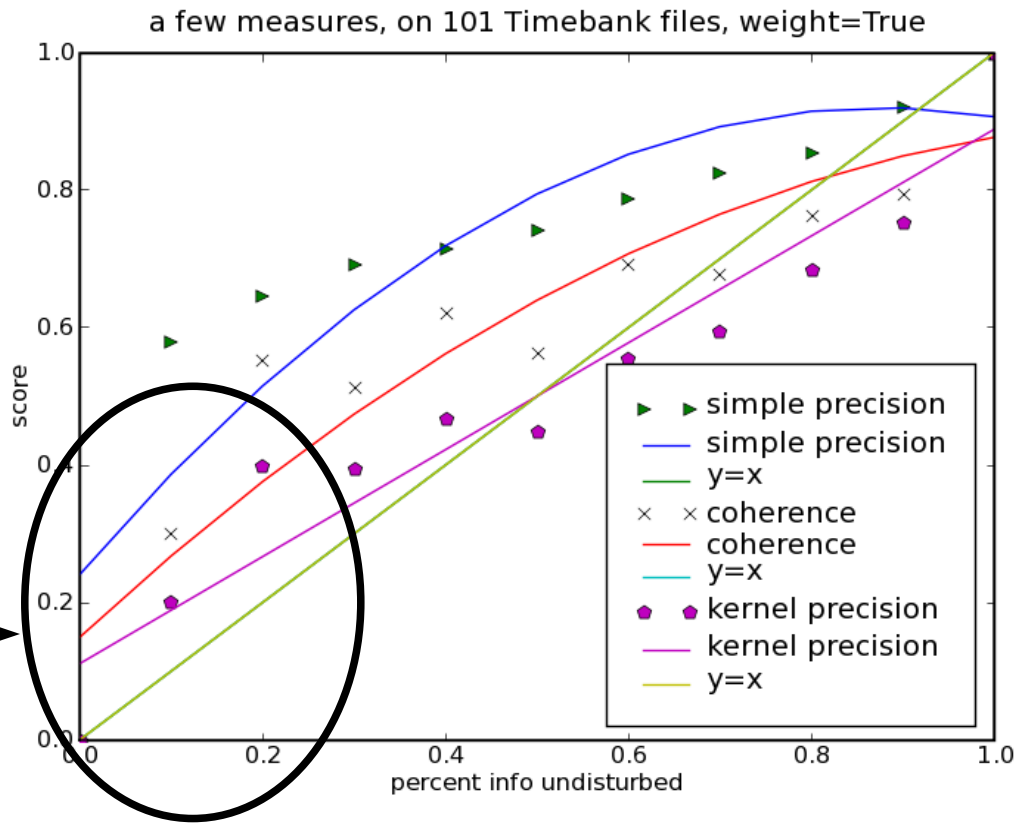
Removing information

- Traditional recall:
  inverted square root
- Kernel recall:
  linear
- Problems with unlinked nodes

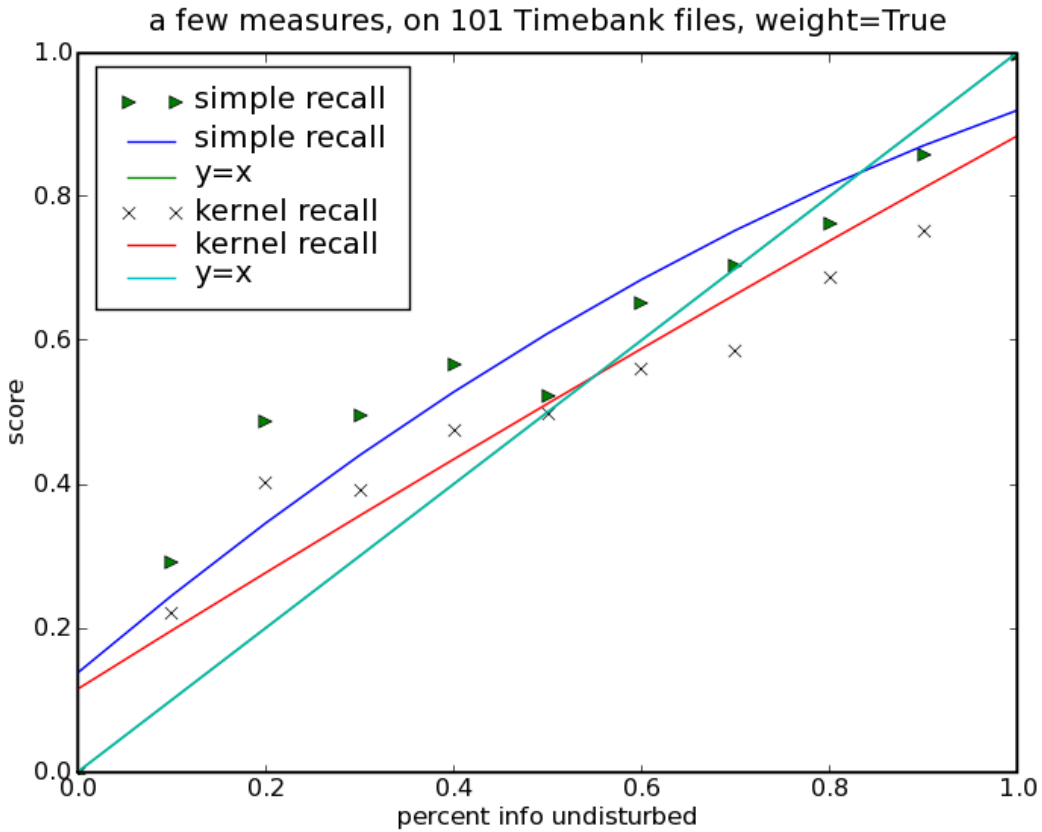


Changing information (precision)

- Same phenomenon
- Change of information must preserve consistency
- Leads to unstable results when very few information



Changing information (recall)



Conclusion

- Traditional metrics does not evolve linearly as they should do

- Treating only "important" relations seems to improve the behaviour of evaluation
- Assumption to be checked : these relations are the one humans consider important in a text
- But:
  - A way to find "important" relations without losing information is still to be found
  - Yet, in an evaluation, "non-important" relations may deserve to be rewarded
  - Our kernel recall/precision do not evaluate properly an annotation

Thank you!