Temporal information in Topic Detection and Tracking

Juha Makkonen

jamakkon@cs.helsinki.fi

University of Helsinki – Department of Computer Science

Topicspotting

Choose no life. Choose research. Choose computer science. Choose a kind of information retrieval that you've never heard of. Choose topic detection and tracking. Choose dead lines. Choose friggin' huge corpus that eats your hard drive. Choose part-of-speech taggers. Choose ontologies. Choose feature selection and ponder why it does not work. Choose weighting-schemes. Choose testing. Choose ‘top-of-my-head’ similarity measures. Choose edit-distances. Choose starting from scratch. Choose yet-another-kind-of-clustering or what-ever that is supposed to do the job. Choose an evil baseline. Choose TFIDF. Choose sitting in front of the monitor drinking coffee, rotting away at the end of it all.

Choose your problems

Choose TDT.

Outline

- About News
- Topic Detection and Tracking
- Similarity (of Hands)
- Temporal Information
- Ongoing work & further ambitions
- Conclusions

About news

- *"News is not what happens, but what someone says has happened or will happen"* (Sigal, 1986)
- *"Journalism, like any other storytelling, is a form of fiction operating out of its own conventions and understandings and within its own set of sociological, ideological, and literary constraits."* (Manoff & Shudson, 1986)
- *"Every newspaper reporter should answer the questions, What? Who? Where? When? Why? and should do it in the first paragraph as nearly as possible."* (Manoff & Shudson, 1986)
- Good news?

What is a event?

- An event is the tension between two states S and S' (Pachter, 1974)
- An event is something non-trivial taking place somewhere at some time. (Yang, 1999)
- *"An event is a specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences."* (Cieri et al. 2002)
- *"A topic is an event or an activity, along with all related events and activities."* (Cieri et al. 2002)
- Event as a narrative (e.g. Falk 1989)

Topic Detection and Tracking

- Event-based information retrieval (Allan, 2002)
- Five tasks:
  - story segmentation
  - first story detection
  - topic tracking
  - topic clustering
  - link detection
- online monitoring of news stream

Problems in TDT

- In text categorization one tries to model the underlying distribution of documents X and classes C

$$\check{h} : X \times C \rightarrow \{-1, 1\}$$

- A hypothesis is built from labeled training samples.
- A hypothesis is evaluated with testing samples.
- Compared to categories, TDT topics are smaller and change over time.

Problems in TDT (2)

- Suppose there is similar distribution of documents X and events E

$$\check{g} : X \times E \rightarrow \{-1, 1\}$$

- The domain E is time-dependent: $E_{train} \neq E_{test}$.
- In fact, we have no a priori knowledge about E_{test}.
- We are left with pair-wise similarity of the documents

$$k : E \times E \rightarrow \{-1, 1\}$$

- Any two documents discussing the same event are ideally similar *in the same way*.

Problems in TDT (3)

- The events evolve which alters the vocabulary.
  - The event model has to address change.
- Allan et al. (2000) showed the inadequacy of the current FSD methods based on tracking.
  - FSD is *"either impossible or requires substantially different approaches"*.
- Pair-wise comparisons require exhaustive computation.
  - The search-space can be confined with traditional IR techniques.

Observations

- News is something new about something known.
- News events evolve.
- News stories have structure and conventions.
- Proper names and locations are particularly valuable.
- Full-text similarity does not pay off.

How to go about similarity of news stories?

Similarity

- What is similarity?
- All conceptual thinking is preceded by similarity.
- For one thing, similarity is a word used by people (to denote so-and-so).
- *“The question "What is a word really?" is analogous to "What is a piece in chess?""* (Wittgenstein, 1953)
- We could say that similarity is a language game, where 'similarity' has “legal” and “illegal” moves.
- We could say the language games are learned as part of culture/being socialized.

Illustrative Example

- What is similarity to a computer? E.g.
  - identity: $A = A$
  - metric: $d(A, B)$
- Asymmetric similarities (e.g. Tversky, 1977)
- Family resemblance: a network of similarity relations linking the members of a class (Wittgenstein, 1953).

Illustrative Example

- Consider hands H_1 and H_2 drawn from two different decks.

suits	H_1	H_2	$H_1 \cap H_2$
$\spadesuit$	2, 3, K	3, 7, Q	{3}
$\diamondsuit$	6, 7, 10	5	
$\heartsuit$	4, 6, 7, A	4, 7, K	{4, 7}
$\clubsuit$	3, 4, 10, J, Q	9	

- How do we determine their similarity?
  - intersection based on identity: $\spadesuit 6 = \spadesuit 6$.
  - similarity based on card correlations (a *game*)

Illustrative Example

- Various card games define equivalence within suits and/or values:
  - Freecell solitaire: ♠6 and ♣6 are interchangeable
  - Poker: straights, flushes (suits equal, high card tie-breaker)
    $\{\spadesuit 6 - \spadesuit 5 - \spadesuit 4 - \spadesuit 3 - \spadesuit 2\} \equiv \{\heartsuit 6 - \heartsuit 5 - \heartsuit 4 - \heartsuit 3 - \heartsuit 2\}$
  - Euchre: one suit is trump; equivalence changes accordingly
- Cards *denote* something, they have a meaning within the game.

Similarity of Hands

- Leap to information retrieval
  - mapping:

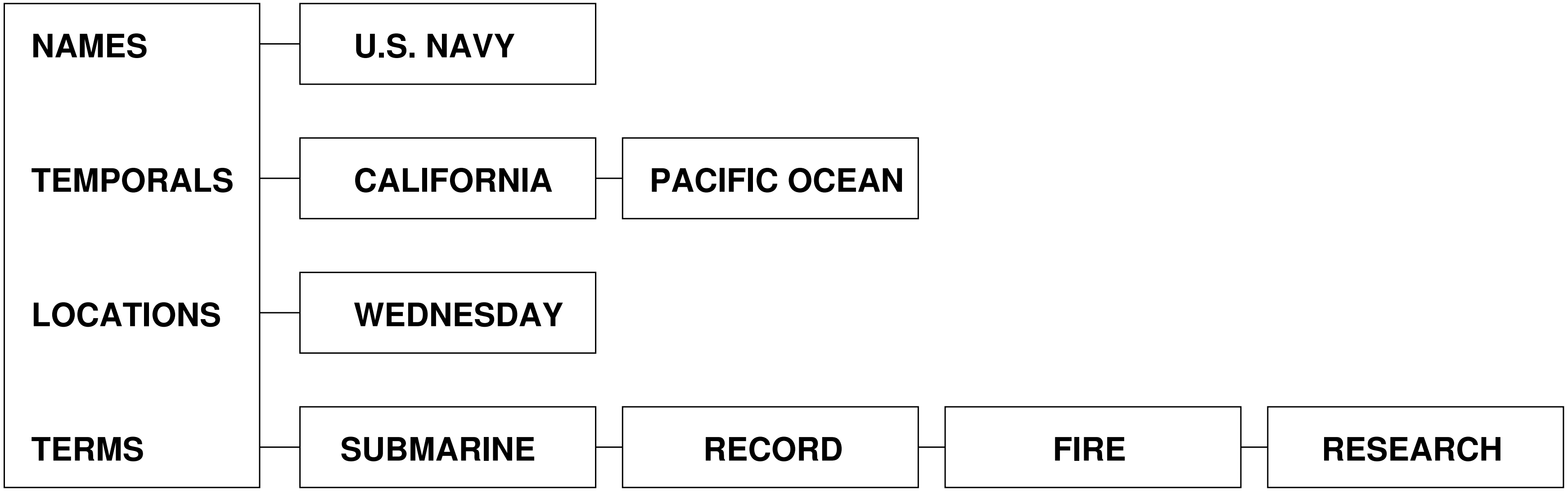
    deck ⇒ feature-space

    card ⇒ term (feature)

    hand ⇒ document
  - Some terms are mutually relevant via their meaning.
    - split the feature-space into classes of "types of meaning"
    - temporal expr. (WHEN), locations (WHERE), proper names (WHO), general terms (WHAT)
    - a designated similarity measure within each *semantic class*
    - naively, if the meaning of a word is in its relation to other words, we are introducing simple semantics.

    suit ⇒ semantic class

Similarity of Hands



"The U.S. Navy diesel research submarine that holds the world's deep-diving record caught fire in the Pacific Ocean off California on Wednesday..." (Washington Post, May 22, 2002)

Similarity of Hands

- The similarity of two multivectors is determined *classwise*, thus each class can have its own similarity measure.
- Such measure can adopt an ontology: a timeline or a geographical hierarchy, for instance.
  - sim(London, Paris) < sim(London, Newcastle)
- The overall similarity of two event vectors is a vector

$$\mathbf{v} = (v_{loc}, v_{name}, v_{temp}, v_{term})$$

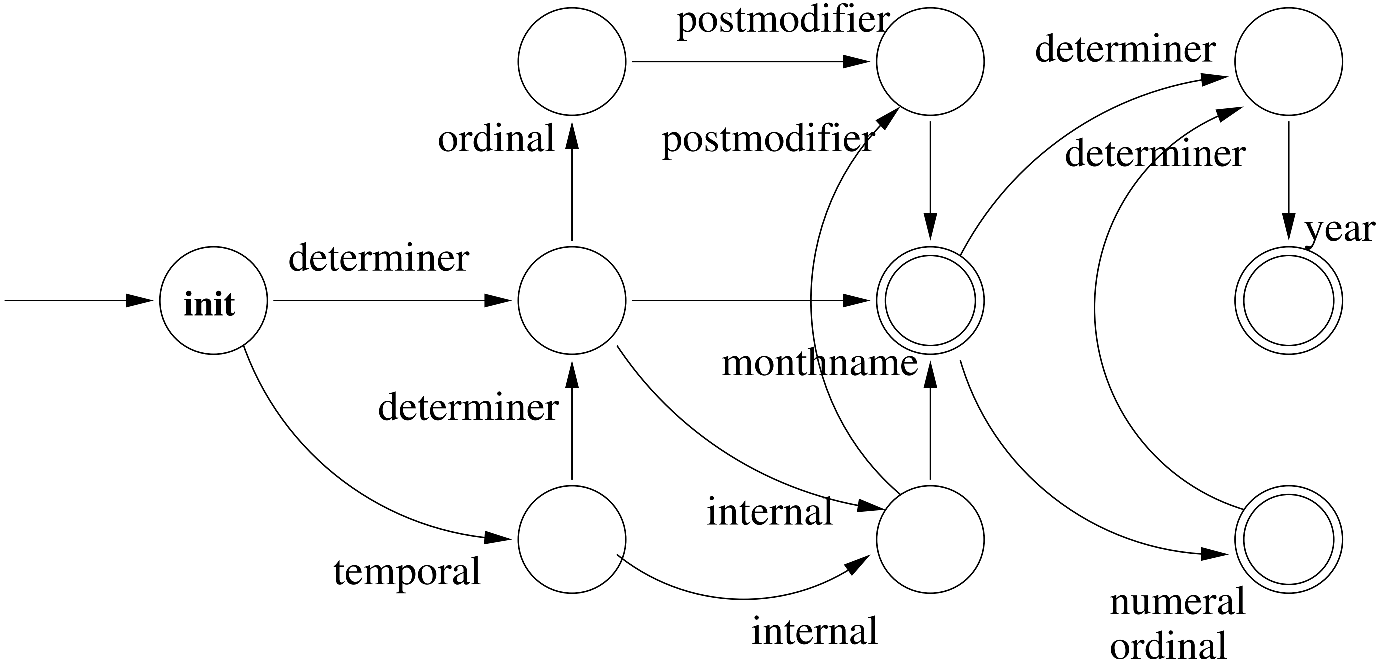
- Learn the 'similarity' from vectors **v**: label comparisons of similar documents with $+1$ and of dissimilar with -1.
- **Similarity of hands** is a 'game' of similarity that is played suitwise.

Temporal Information

- An expression can be
  - explicit: *"the 19th of August 2003"*,
  - implicit: *"today"*, *"Tuesday afternoon"*, or
  - vague: *"since April"*, *"two years ago"* .
- For example
  "The winter of 1974 was cold. The next winter will be colder."
  "The winter of 1974 was cold. The next winter was colder."
- Resolving the meaning of the latter winter requires
  - the *reference time* or the *utterance time* and
  - the tense of the relevant verb.

Temporal Information

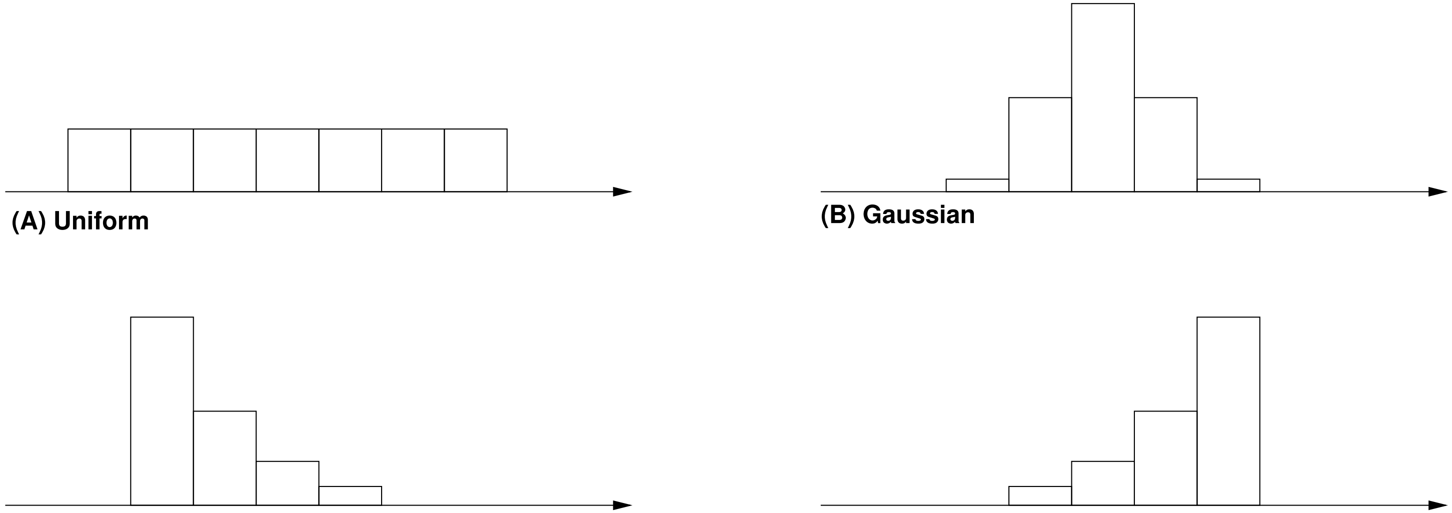
- The terms are used to build automata (Makkonen and Ahonen-Myka 2003).



“The strike started **on the 15th of May 1919**. *It lasted* **until the end of June**, *although there was still turmoil* **in late January next year***”*.

Temporal Information

- The expression is turned into shift and span functions.
- We map the expressions via a *calendar* (timeline, granularities, conversion functions)
- The expressions are mapped as intervals $[t_{start}, t_{end}]$ of the bottom granularity which in our case is *day*.
- Vagueness is introduced via relevance distributions.



(A) Uniform

(B) Gaussian

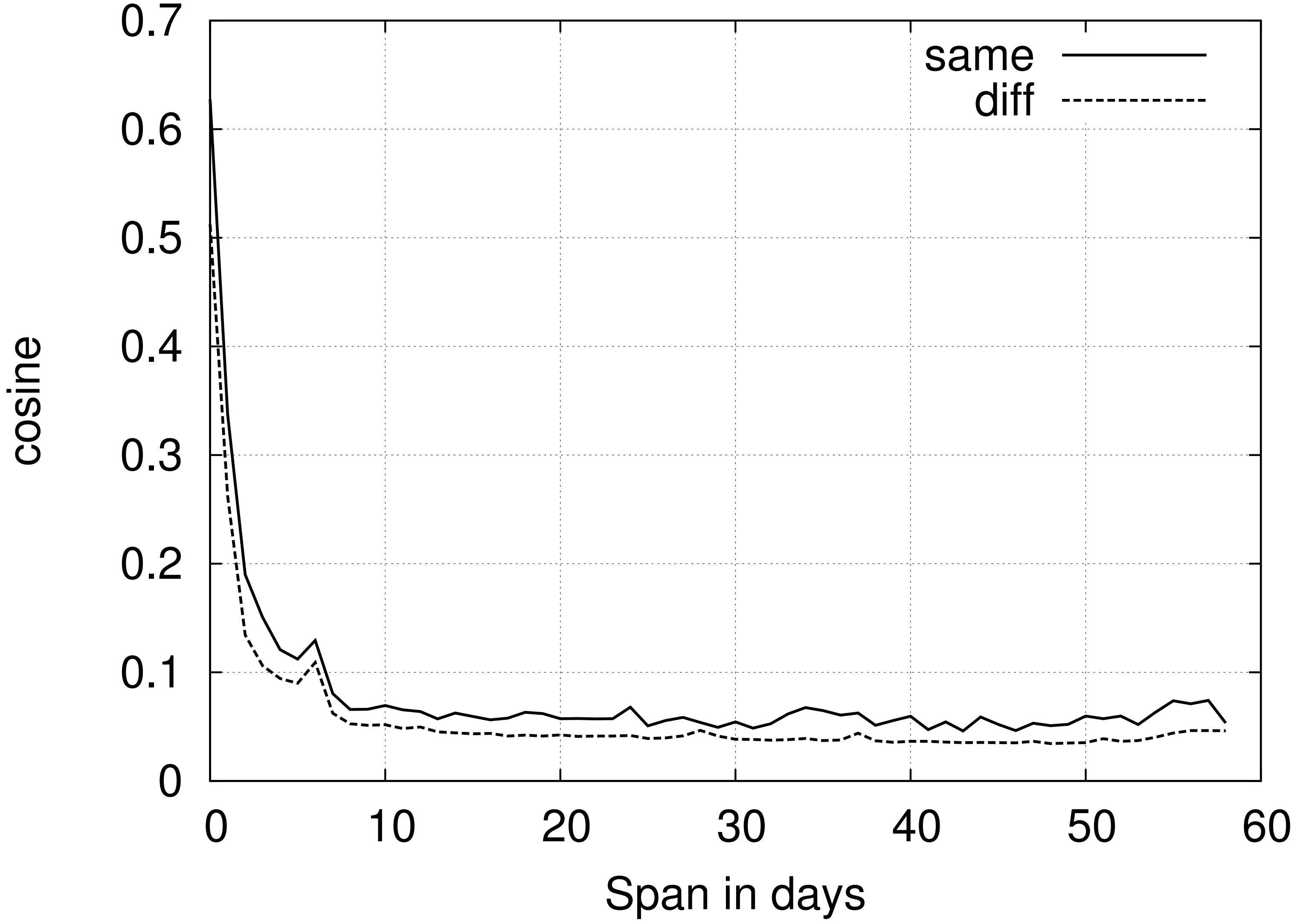
(C) Exponential

(D) Reversed Exponential

Temporal Information

- We are not reasoning about chronological order of news events, but *the similarity of two sets of temporal expressions*.
- The set of temporal expressions is mapped onto a timeline.
- Similarity of two sets is based on overlap, cosine, etc.
- NOTE: the news stories published on the same day have typically high temporal similarity.

Temporal Information



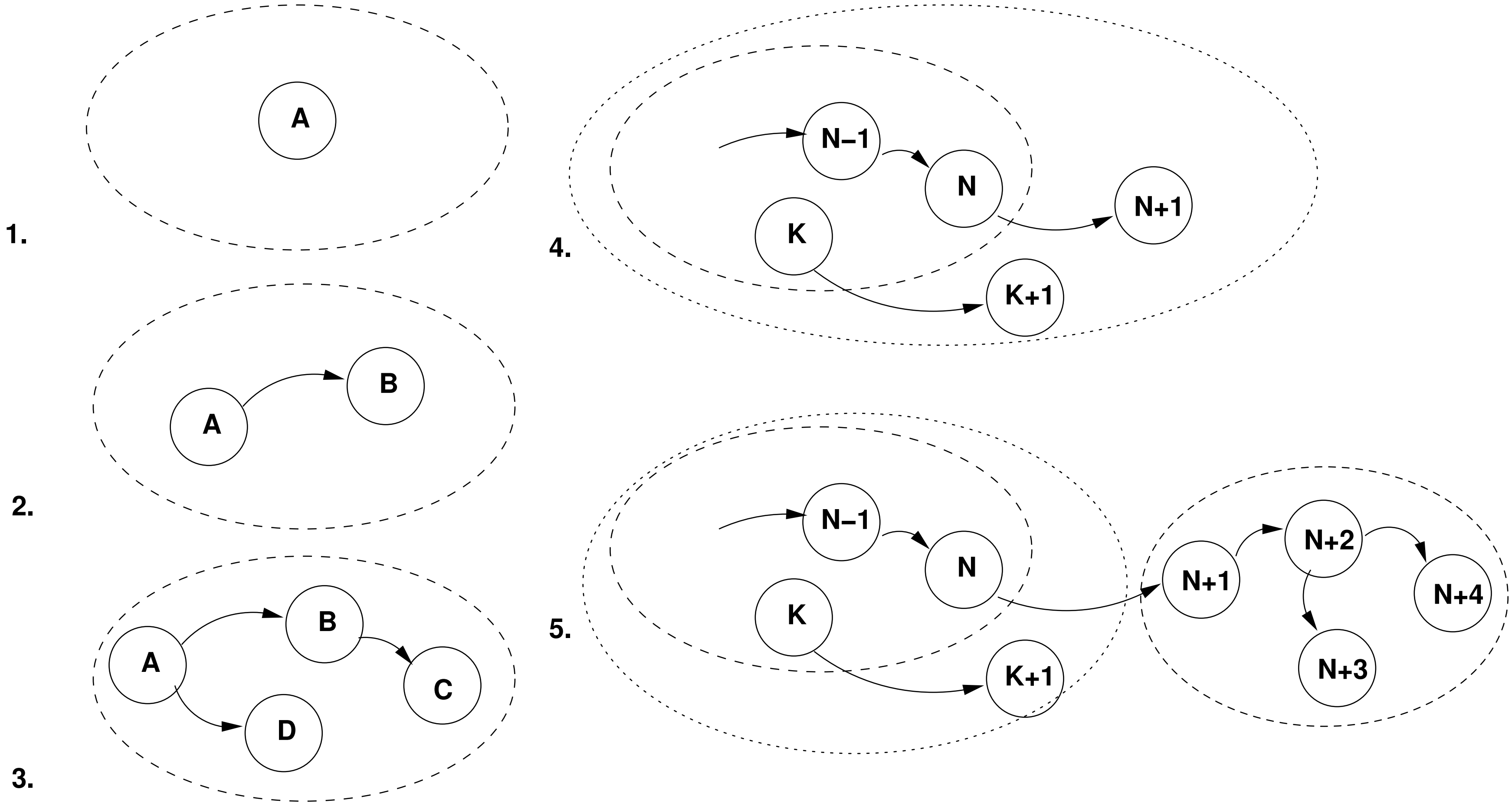
Ongoing work

- Tuning of the distributions for temporal information
- Reasonable similarity measure for locations
  - geographical relevance of locations
  - statistical relevance of location terms
- Experiments

Further ambitions

- Should this approach work we could
  - examine the behaviour of confirmation and new information
  - try to address the evolution of events
  - maybe apply the approach to other event-based data (historical accounts, annales, etc.)
  - similarity of two general terms via thesaurus, WordNet, etc. ?
- further analysis of
  - what constitutes an event
  - event as a narrative? (Makkonen, 2003)

Further ambitions



Conclusions

- News data is produced by reporters and has conventions, structure etc.
- TDT is a form of event-based information organization.
- Similarity of hands
  - splitting of feature-space into semantic classes
  - within a class the similarity can be what-ever
  - similarity is determined class-wise
- Temporal expressions
  - Recognition from text
  - Relevance distributions on a timeline.
  - Similarity of two sets of expressions.

References

- Allan J, (2002) Topic Detection and Tracking. Kluwer, Norvell (MA).
- Allan J, Lavrenko V and Jin H (2000) First story detection in TDT is hard. Proc. ACM CIKM 2000, 374–381.
- Bell A (1999) News stories as narratives. The Discourse Reader, Routledge, London, UK, 231–251.
- Cieri C et. al (2002) Corpora for topic detection and tracking. In Allan (2002), 33–66.
- Falk P (1989) The Past to Come. Economy and Society, 17(3): 374–394.
- Makkonen J and Ahonen-Myka H (2003) Utilizing Temporal Expressions in Topic Detection and Tracking. Proc. ECDL 2003, 393–404.
- Makkonen J (2003) Investigations on Event Evolution in TDT. Proc. HLT-NAACL 2003 Student Workshop, 43–48.
- Manoff R K and Schudson M (1986) Reading the News. Pantheon Books, New York.
- Pachter H (1974) Defining an event. Social Research 41(3):439–466.
- Sigal L V (1986) Sources of News. In Manoff and Schudson (1986), 9–37.
- Tversky A (1977) Features of Similarity. Psychological Review, Vol. 84(4): 327–352.
- Wittgenstein L (1953) Philosophische Untersuchungen. Translation Filosofisia tutkimuksia, WSOY, Juva, 1996.