Motivation

Distributional Hypothesis

Words that occur in the same contexts tend to have similar meanings

(Zellig Harris, 1954)

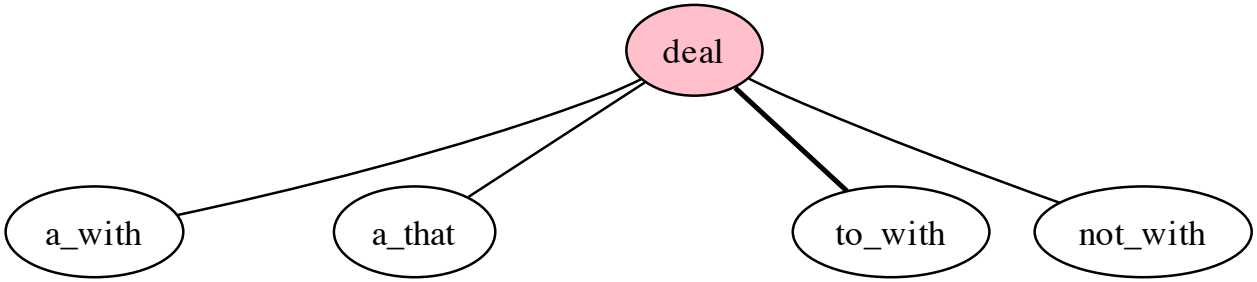
We will leverage this in a translation setting:

- Use the contexts to cluster translation units into groups
- Units in the same group expected to be semantically and syntactically similar
- Then use these cluster labels to guide translation
  - lexical selection: translating ambiguous source word/s
  - reordering: consistent syntactic patterns of reordering

Monolingual Example

Task: cluster words into their parts-of-speech.

Illustrate by starting with the word ‘deal’ (noun or verb):

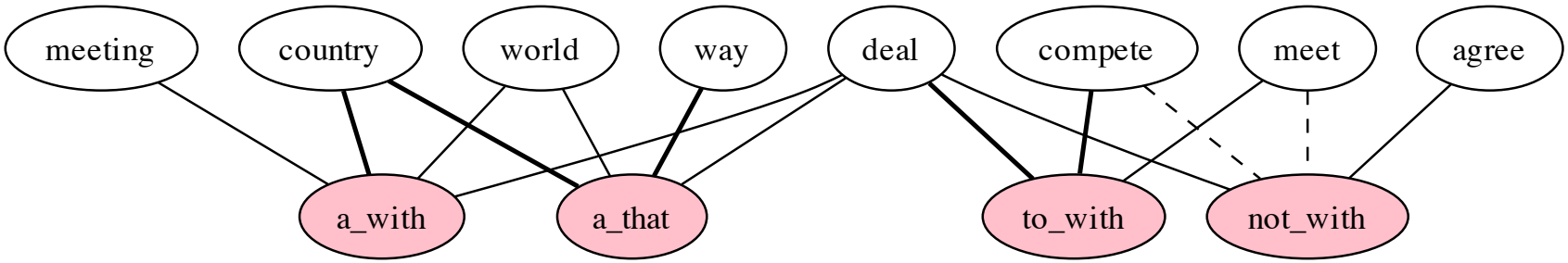


Step 1: Find contexts for ‘deal’

Monolingual Example

Task: cluster words into their parts-of-speech.

Illustrate by starting with the word ‘deal’ (noun or verb):

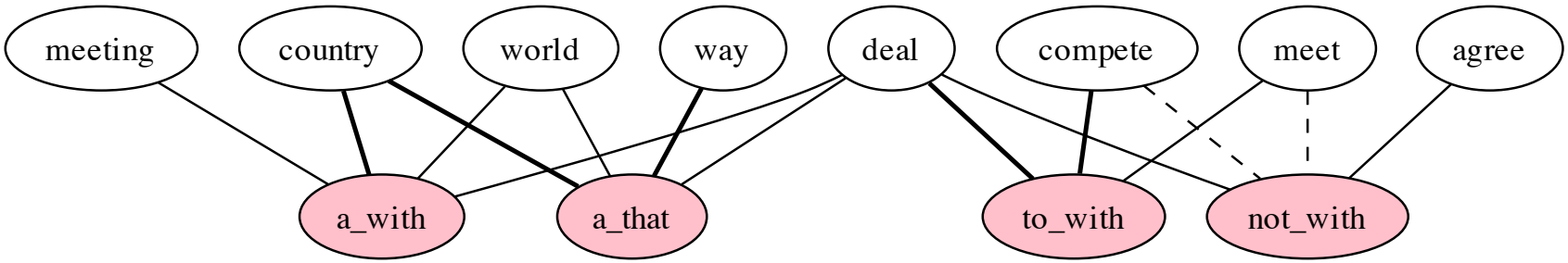


Step 2: Find other words which occur in these contexts

Monolingual Example

Task: cluster words into their parts-of-speech.

Illustrate by starting with the word ‘deal’ (noun or verb):



Step 2: Find other words which occur in these contexts

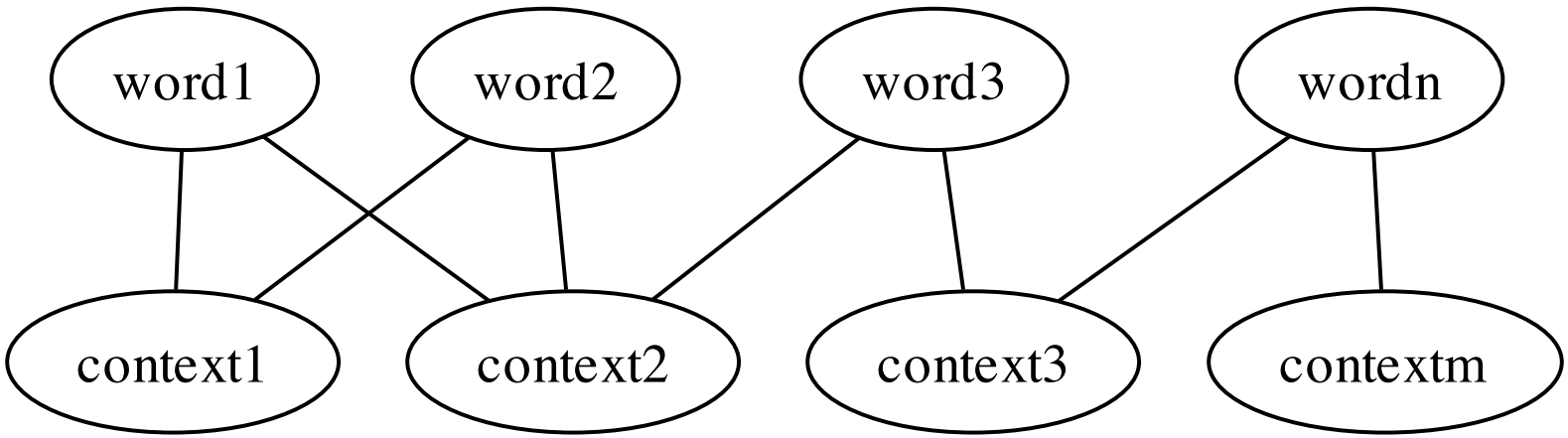
Notice that the instances of deal can be split into two connected sub-graphs:

- noun: the left two contexts “a . . . with” and “a . . . that”
- verb: the right two contexts “to . . . with” and “not . . . with”
- neighbouring words of these contexts share the same PoS

More Formally

Construct a bipartite graph

- Nodes on the top layer denote word types (bilingual phrase pairs)
- Nodes on the bottom layer denote context types (monlingual/bilingual words)
- Edges connect words and their contexts

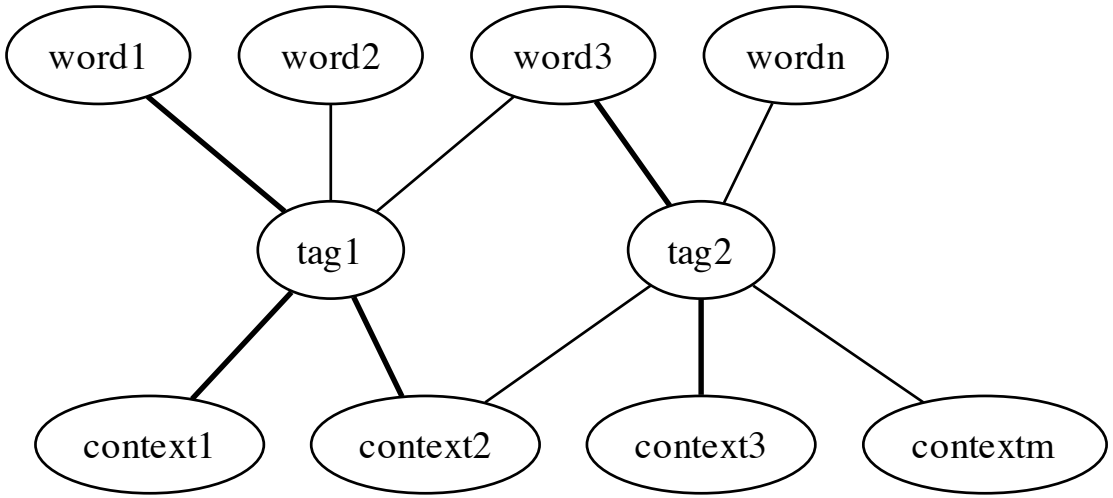


Clustering

Task is to cluster the graph into sub-graphs. Nodes in the sub-graphs should be

- strongly connected to one another
- weakly connected to nodes outside the sub-graph
- could formulate as either *hard* or *soft* clustering

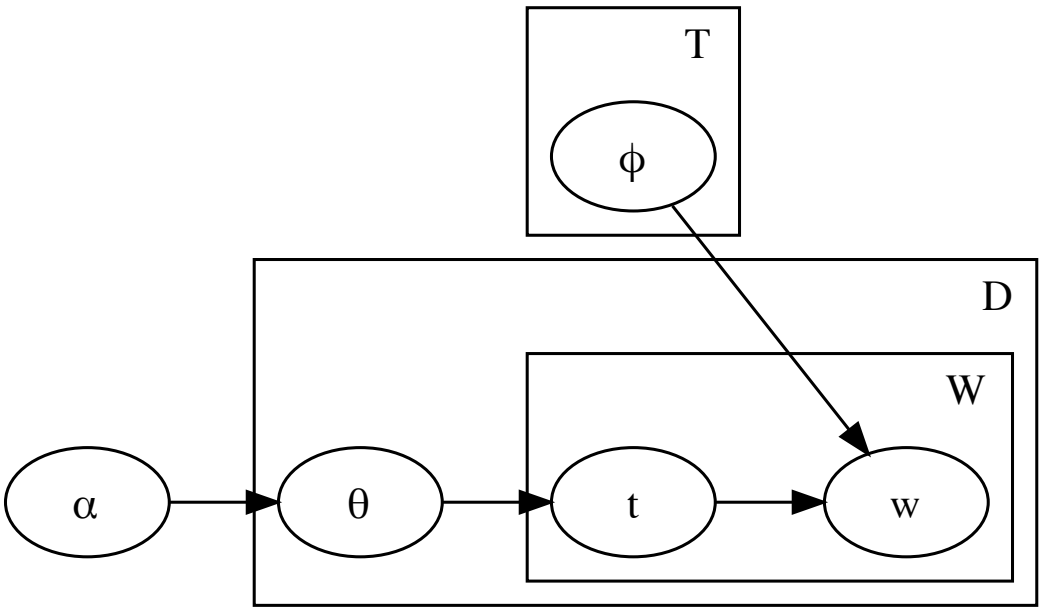
Choose soft clustering to allow for syntactic and semantic ambiguity



Latent Dirichlet Allocation (LDA)

LDA is a generative model which treats documents as bags of words

- each word is assign a topic (cluster tag)
- words are generated from a topic-specific multinomial
- topics are tied across a document using a Dirichlet prior
- $\alpha < 1$ biases towards sparse distributions, i.e., topic reuse
- inferred θ_d describes a document and ϕ_t describes a topic



LDA over Contexts

Generative story:

- for each word type w
- for each of the L contexts
- first we draw a topic t, then generate the context $\vec{c}$ given the topic
- the Dirichlet prior ties the topics for each w
- we're primarily interested in the learnt θ values