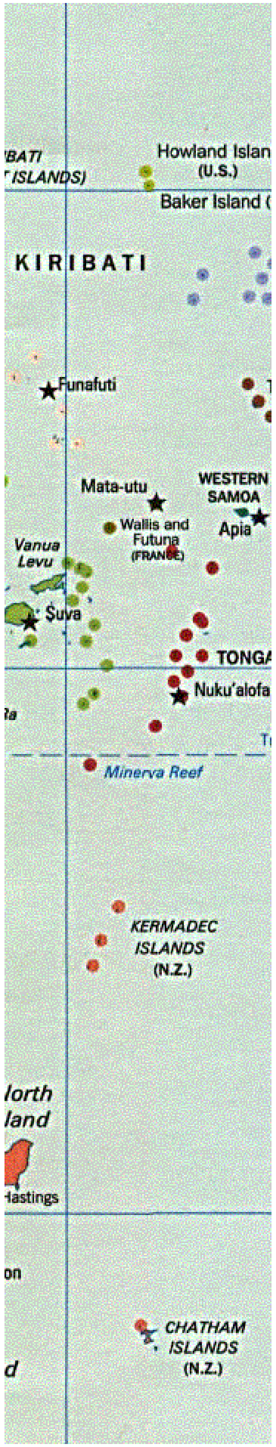


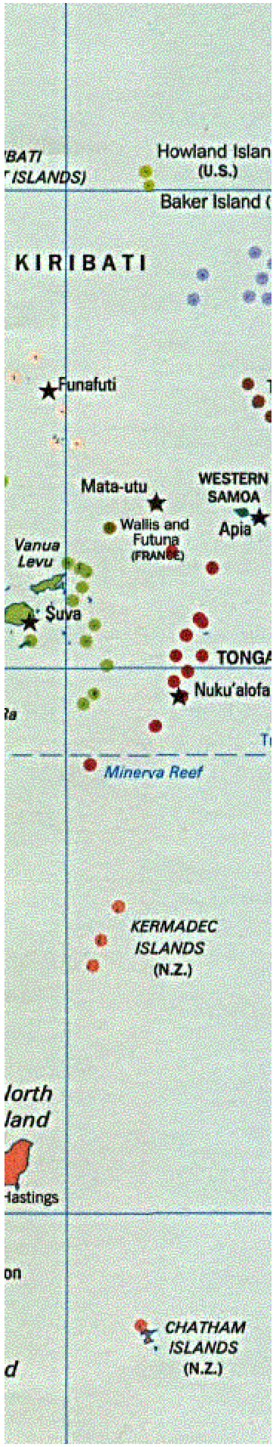
Re-skewing your distribution

Linguistically motivated Sample selection for
Coreference Resolution

Olga Uryupina

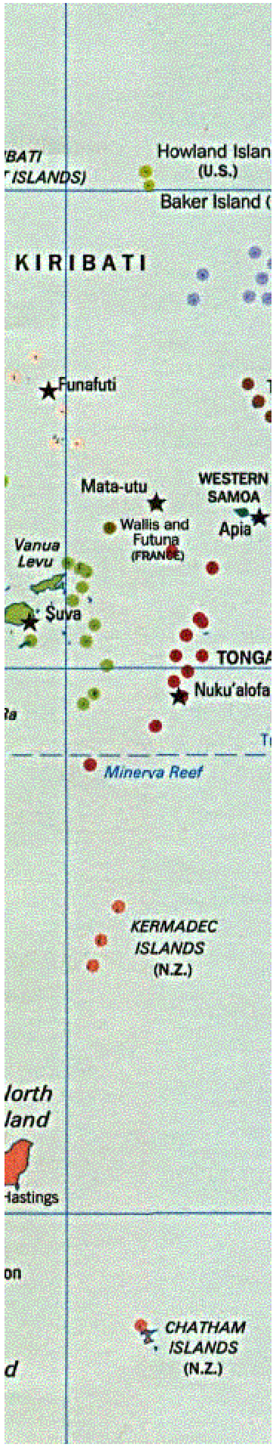
17.06.04

UNIVERSITÄT
DES
SAARLANDES



Overview

- Intro: Machine Learning for CR
- Problem: Generating Training Instances
- Sampling Solutions
- Preliminary Evaluation
- Conclusion and Future Work

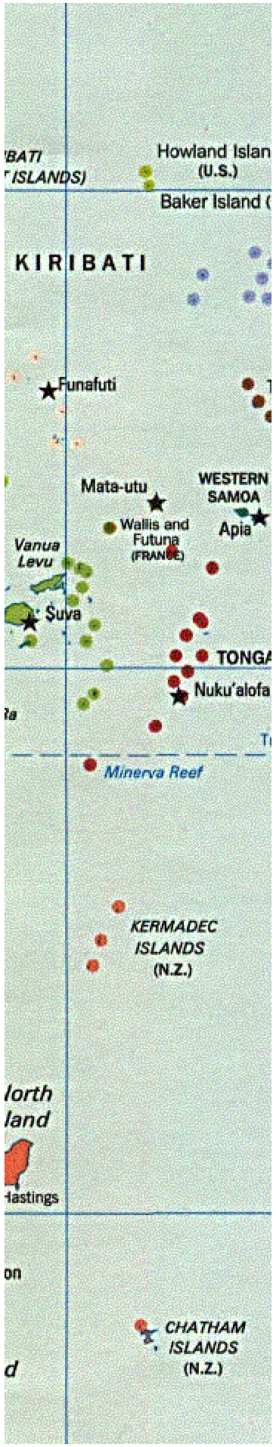


Machine Learning for CR

Coreference chains:

This deal means that Bernard Schwarz can focus most of his time on Globalstar.." said Robert Kaimovitz, a satellite communication analyst at Unterberg Harris in New York. [..] Schwartz said Monday that [..]

UNIVERSITÄT DES SAARLANDES



Machine Learning for CR

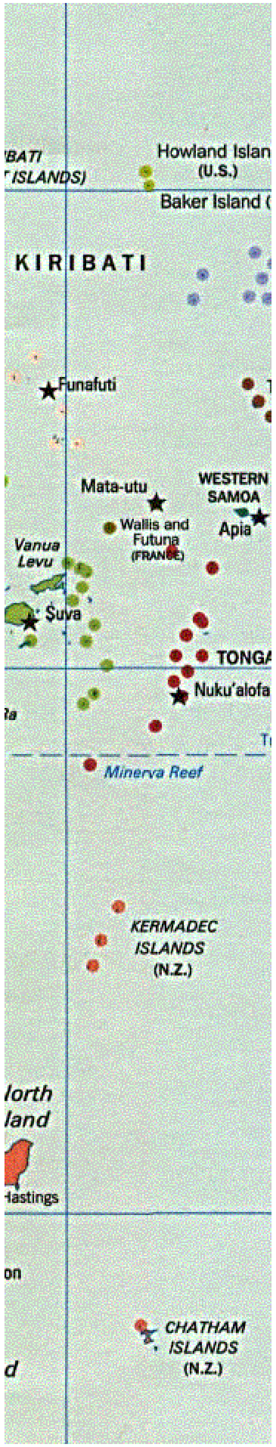
Coreference chains

C1: Bernard Schwarz, his, Schwartz

C2: Robert Kaimovitz, a satellite communication analyst at Unterberg Harris in New York

Machine Learning

Classifier: feature_vector -> class



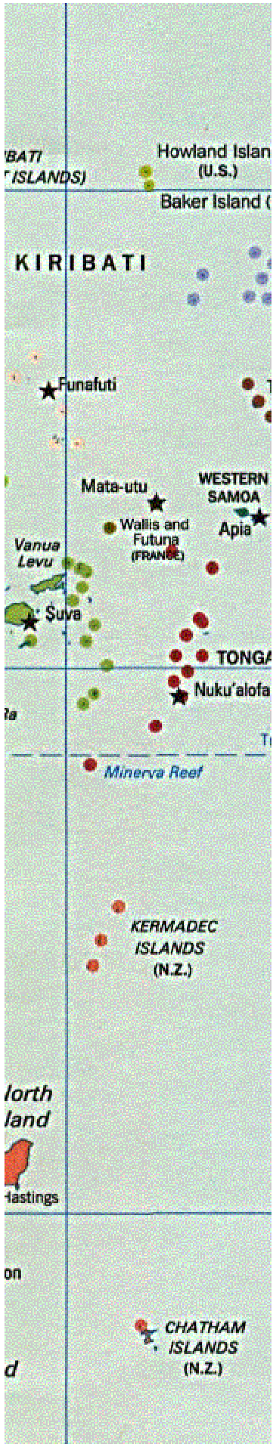
Machine Learning for CR

2-steps approach:

1. Classification – identify [+coreferent] pairs
2. Clustering – merge pairs into chains

Preprocessing:

decompose chains into pairs



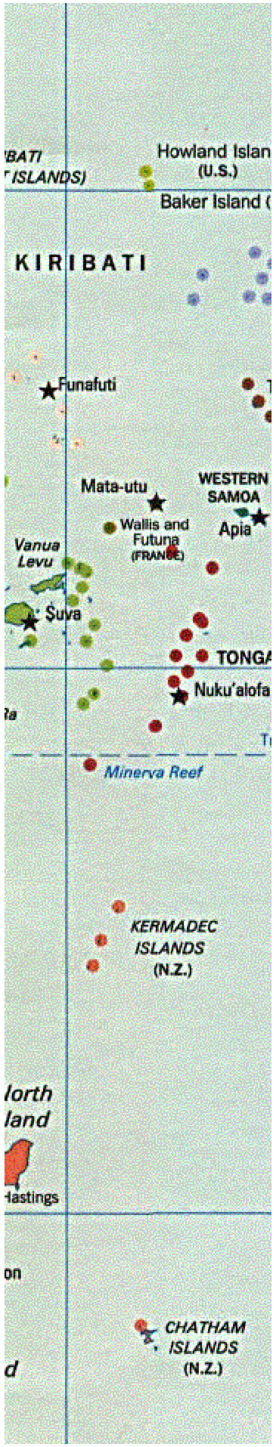
Machine Learning for CR

2-steps approach:

1. Classification – identify [+coreferent] pairs
2. Clustering – merge pairs into chains

Preprocessing:

decompose chains into pairs



Machine Learning for CR

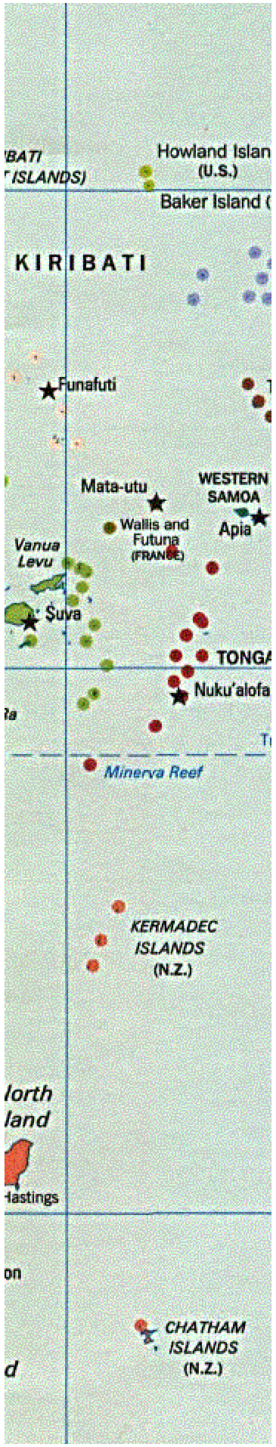
2-steps approach:

1. Classification – identify [+coreferent] pairs
2. Clustering – merge pairs into chains

Preprocessing:

decompose chains into pairs

UNIVERSITÄT DES SAARLANDES



Machine Learning for CR

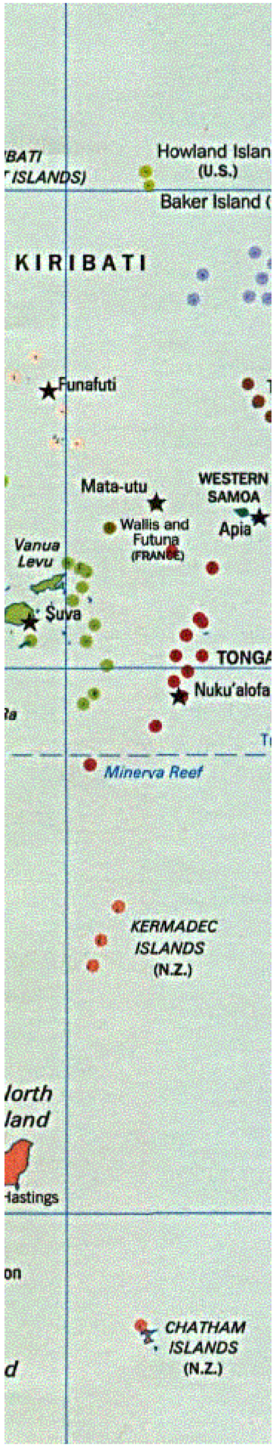
2-steps approach:

1. Classification – identify [+coreferent] pairs
2. Clustering – merge pairs into chains

Preprocessing:

decompose chains into pairs

UNIVERSITÄT DES SAARLANDES

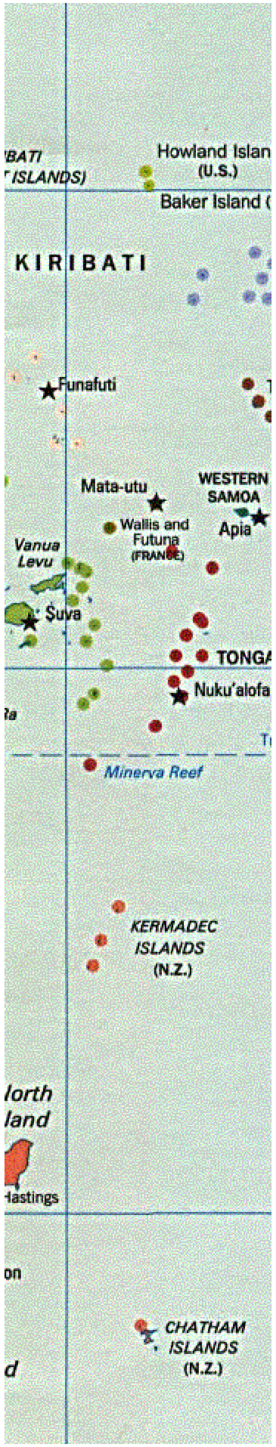


Generating training instances

Standard algorithm

1. Take a markable (anaphor)
2. Pair it with all the preceding ones (candidate antecedent)
3. Assign [±corefernt] class mark
4. Proceed to the next markable

UNIVERSITÄT DES SAARLANDES

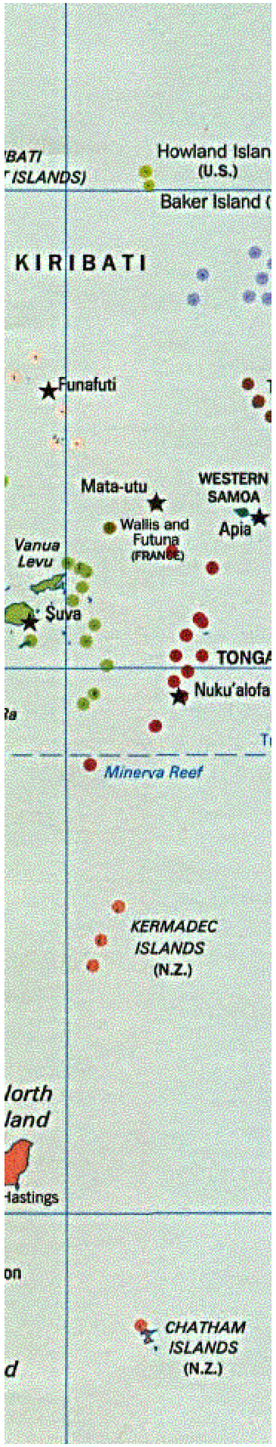


Generating training instances

Back to our example..

This deal means that Bernard Schwarz can focus most of his time on Globalstar.." said Robert Kaimovitz, a satellite communication analyst at Unterberg Harris in New York. [..] Schwartz said Monday that [..]

UNIVERSITÄT DES SAARLANDES



Generating training instances

Back to our example..

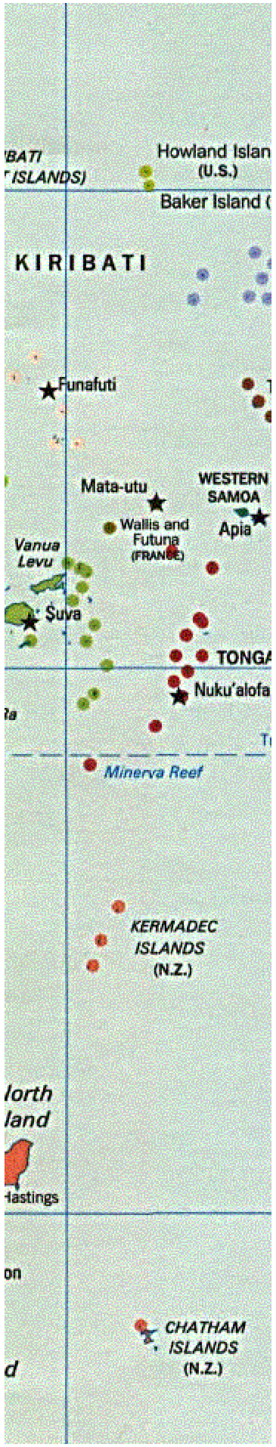
11 markables -> 55 pairs

51 negative pair (This deal, Monday)..

4 positive pairs:

- (Bernard Schwarz, his)
- (Bernard Schwarz, Schwartz)
- (his, Schwartz)
- (Robert Kaimovitz, a sat. comm. analyst)

UNIVERSITÄT DES SAARLANDES



Generating training instances

Problems:

1. Too many negative examples

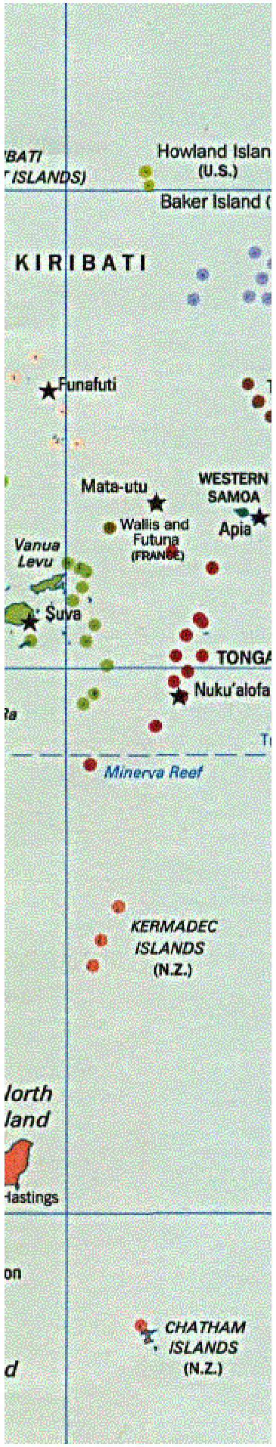
   93% in the toy sample,

   99% in MUC-7

2. Too hard/irrelevant positive examples

   (his, Schwartz)

UNIVERSITÄT DES SAARLANDES



Machine Learning for CR

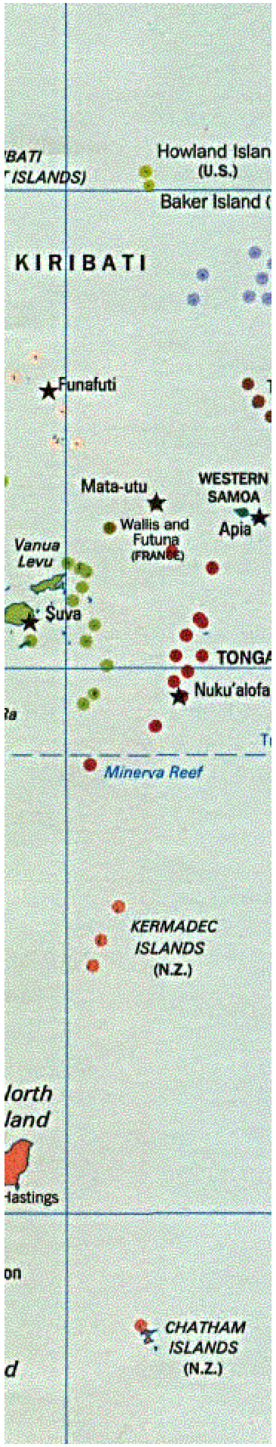
2-steps approach:

1. Classification – identify [+coreferent] pairs
2. Clustering – merge pairs into chains

Preprocessing:

decompose chains into pairs

UNIVERSITÄT DES SAARLANDES

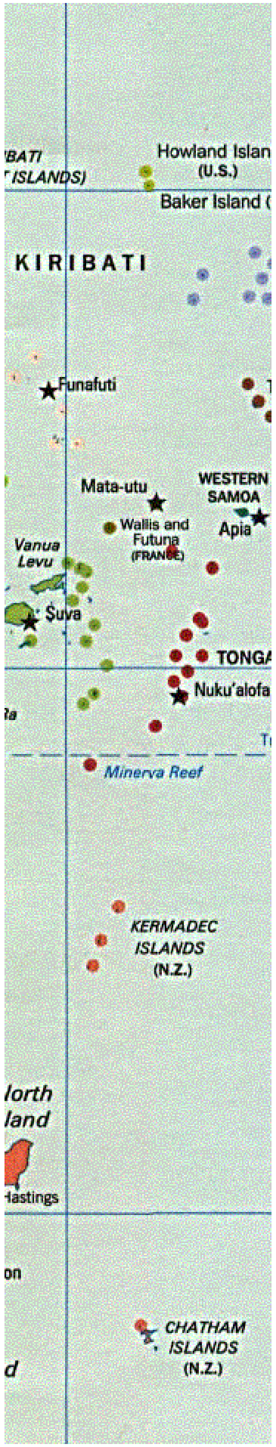


Sampling

Main idea: look at the clustering component and discard unnecessary training items

Expected result: the classifier may get worse, but the overall performance (on chains) increases.

UNIVERSITÄT DES SAARLANDES



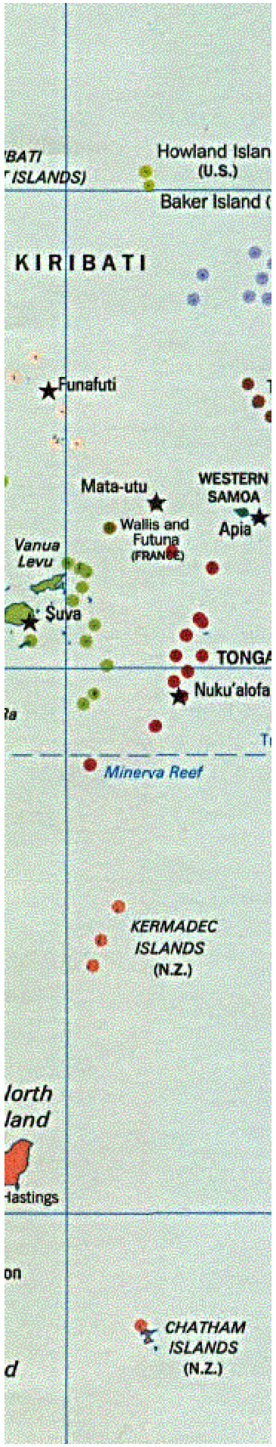
Sampling

Single-link clustering

1. Take a markable (anaphor)
2. Proceed backward, take a markable (antecedent), make a pair (ante, anaph)
3. Submit the pair to the classifier

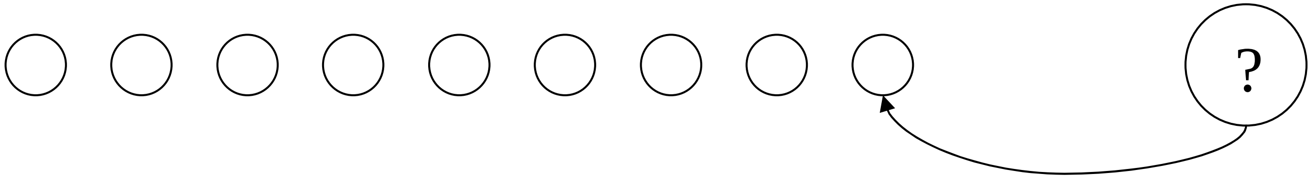
   [+] -> link the anaphor to the antecedent's chain, proceed to the next anaphor

   [-] -> go to step 2

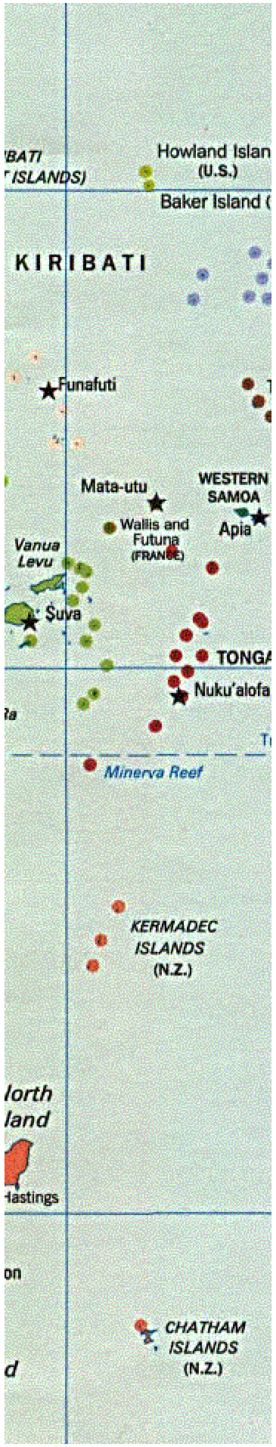


Sampling

Single-link clustering

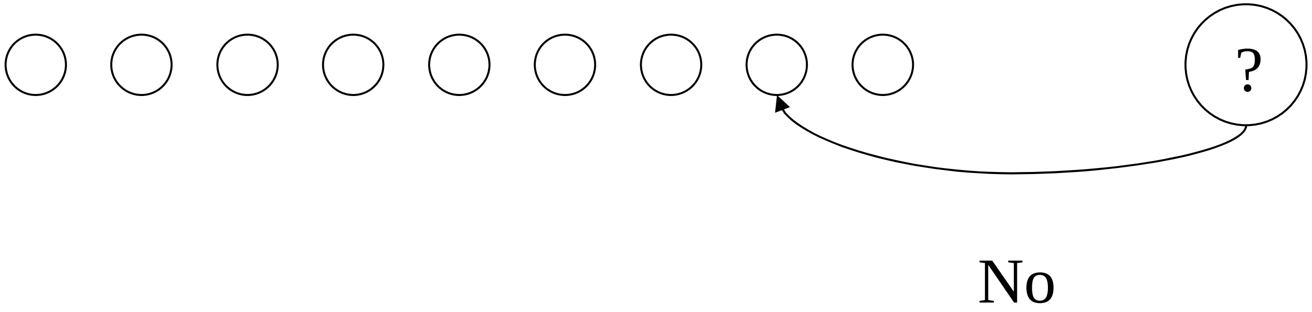


No

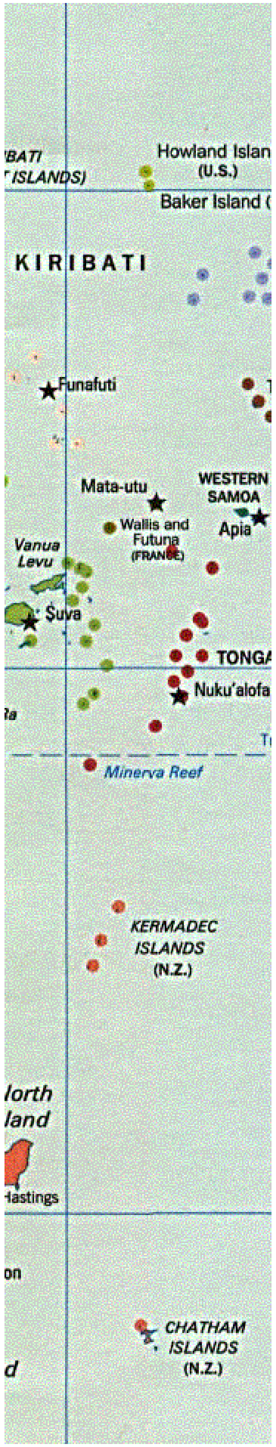


Sampling

Single-link clustering

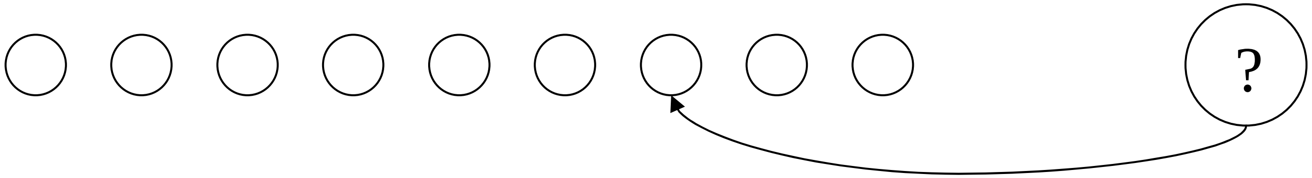


UNIVERSITÄT DES SAARLANDES



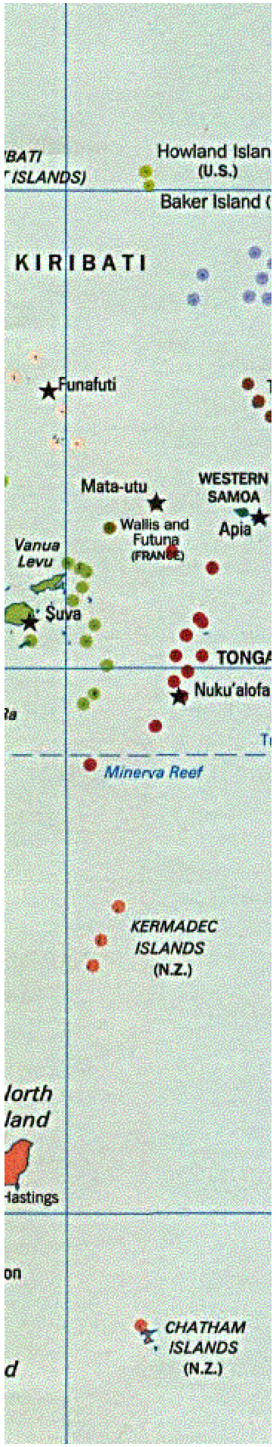
Sampling

Single-link clustering



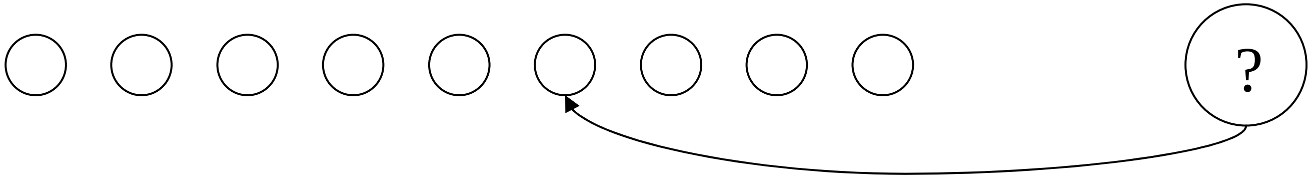
No

UNIVERSITÄT DES SAARLANDES



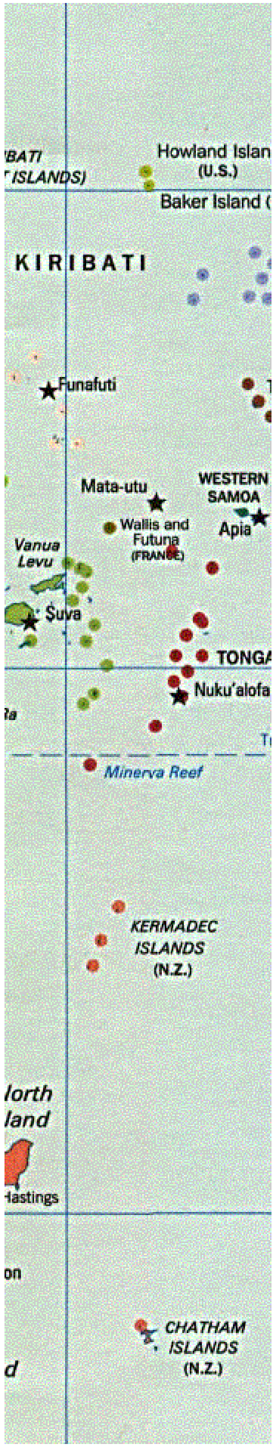
Sampling

Single-link clustering



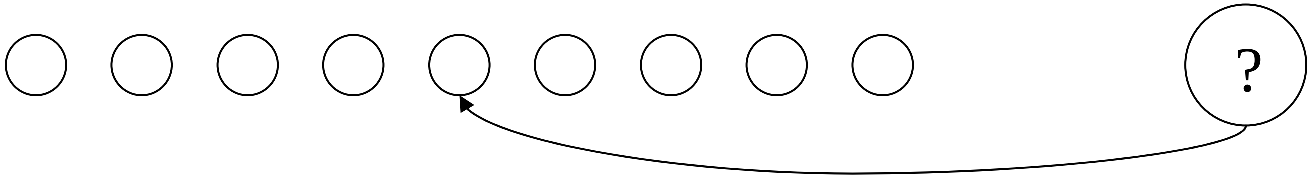
No

UNIVERSITÄT DES SAARLANDES

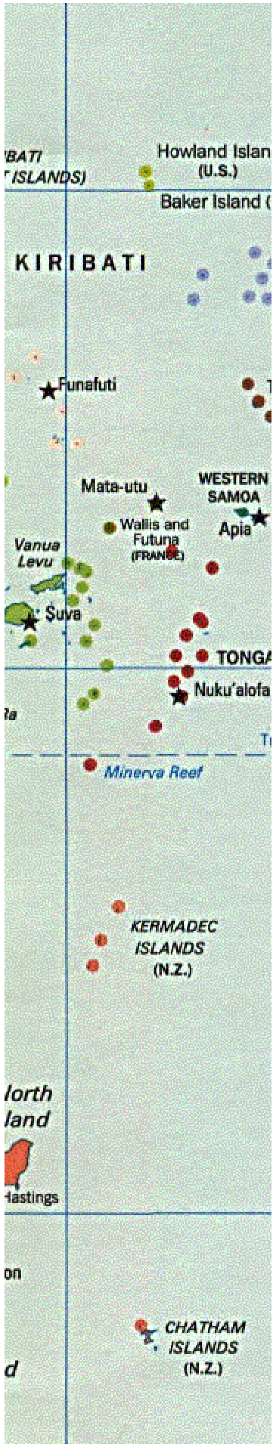


Sampling

Single-link clustering

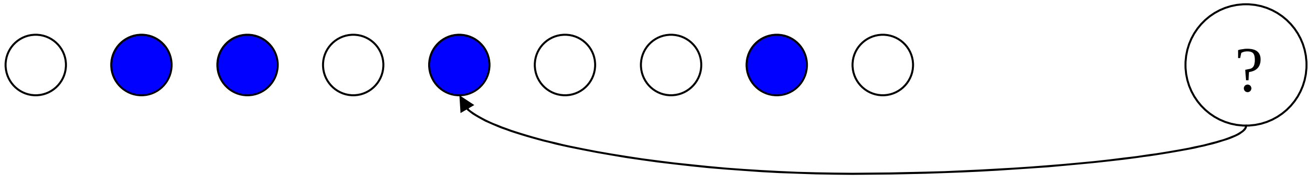


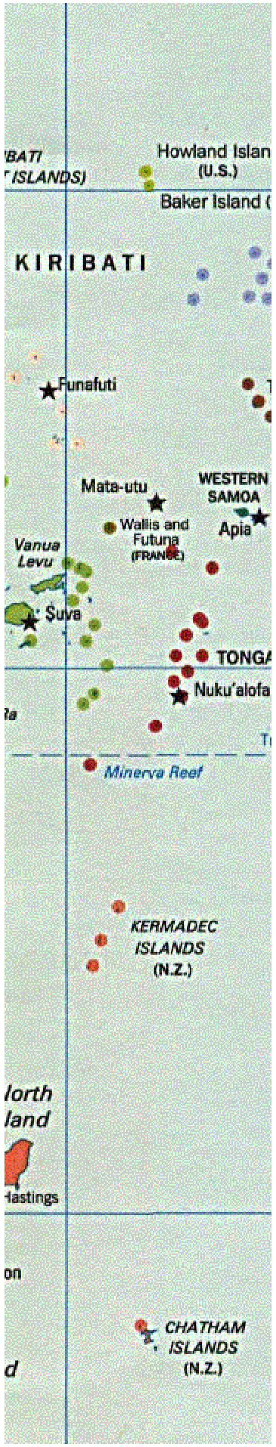
Yes!



Sampling

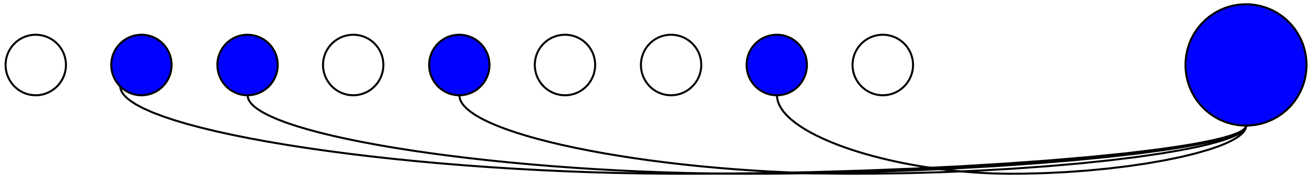
Single-link clustering





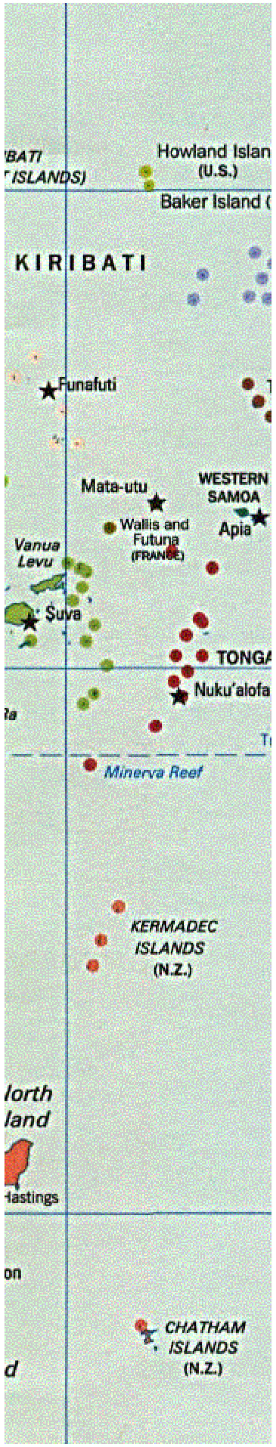
Sampling

Single-link clustering



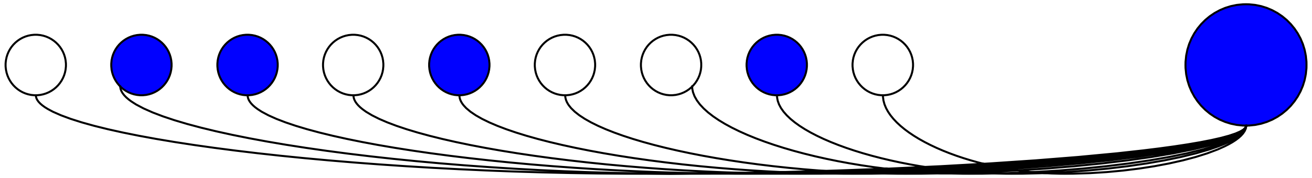
Important properties:

1. Once an antecedent is found, the preceding markables are not processed.
2. Enforces equivalence

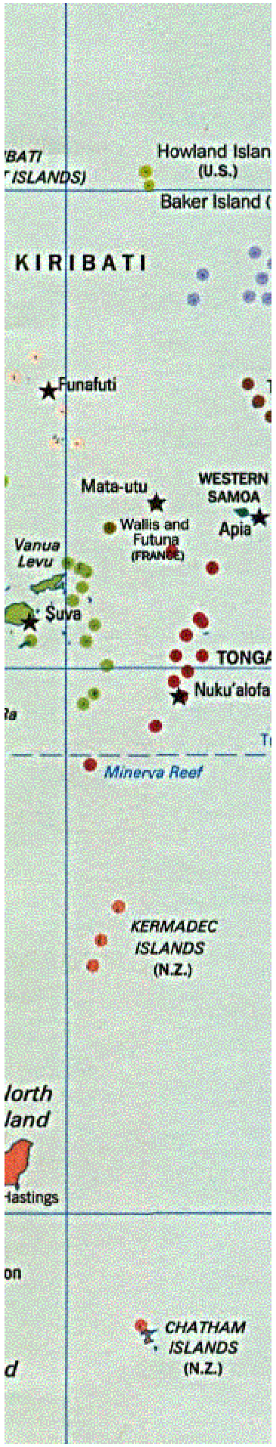


Negative Sample Selection (Soon et al., 2001)

Training data

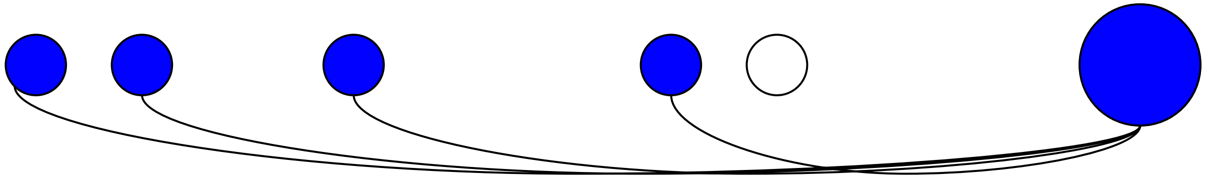


Idea: discard all the negative instances with the candidate antecedents to the left of the rightmost true antecedent

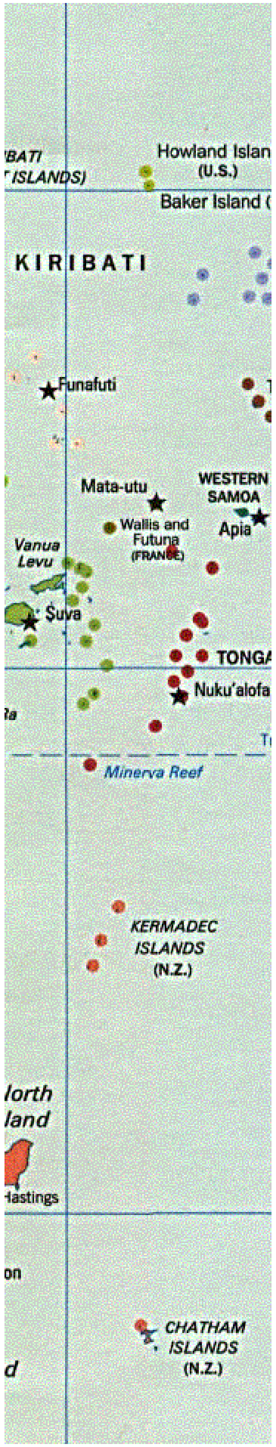


Negative Sample Selection
(Soon et al., 2001)

Training data



Idea: discard all the negative instances with the candidate antecedents to the left of the rightmost true antecedent



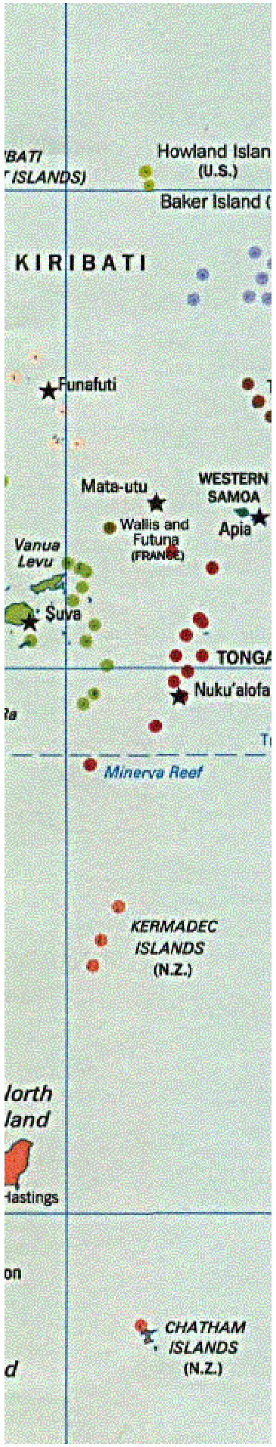
Positive Sample Selection

(Harabagiu et al., 2000), (Ng and Cardie, 2002)

Corpus-based approaches

Idea: identify the easiest positive examples, using various corpus statistics

UNIVERSITÄT DES SAARLANDES

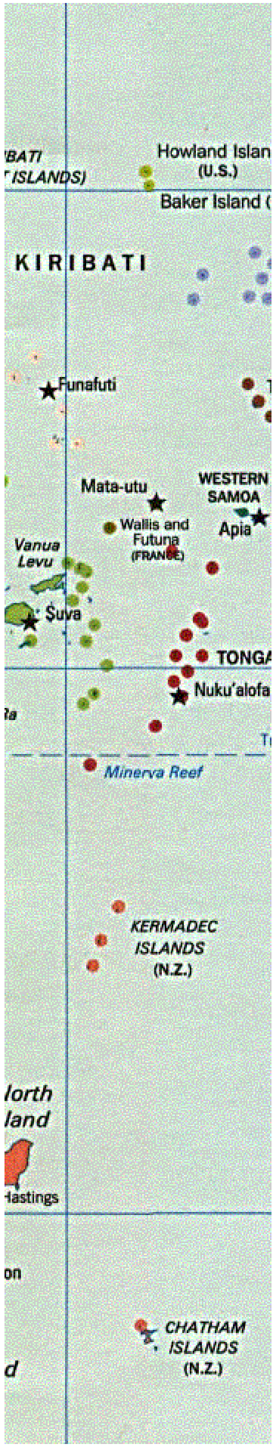


Sample Selection

Linguistically motivated approach

Idea: identify the most relevant positive examples, using linguistic information

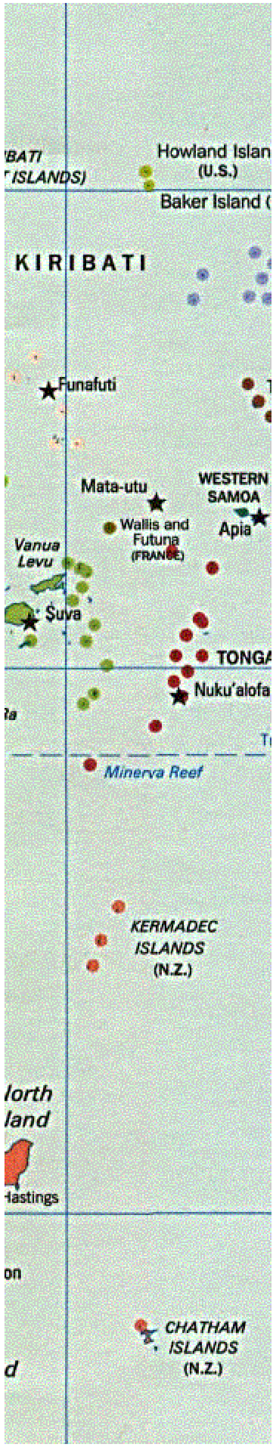
UNIVERSITÄT DES SAARLANDES



Sample Selection

Types of markables:

1. Pronoun
2. Definite Description
3. Proper Name
4. Other (indefinite, bare plural, parsing mistake, NP with a determiner)



Sample Selection

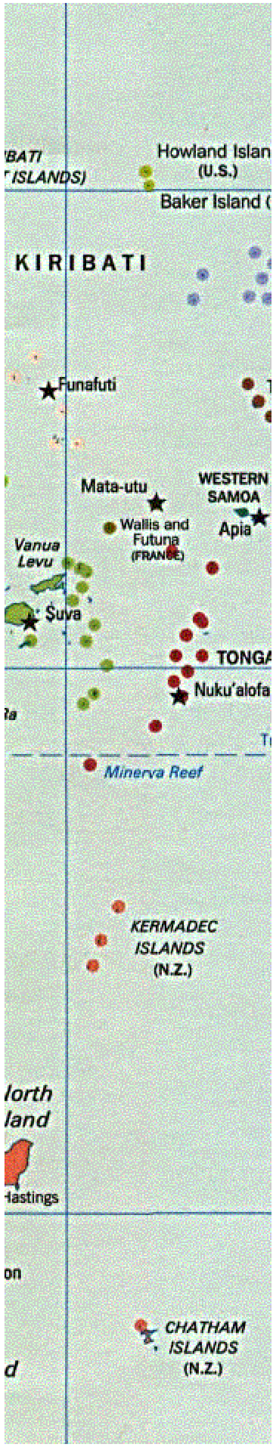
They are really very different..

1. Pronoun

   discourse structure (salience, accessibility..), few preceding sentences

2. Definite Description
3. Proper Name
4. Other (indefinite, bare plural, parsing mistake, NP with a determiner)

UNIVERSITÄT DES SAARLANDES



Sample Selection

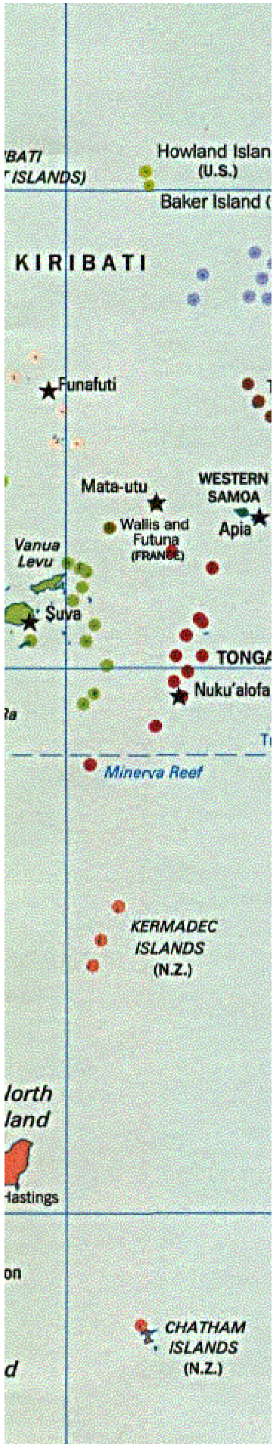
They are really very different..

1. Pronoun
2. Definite Description

   semantic info for head nouns

3. Proper Name
4. Other (indefinite, bare plural, parsing mistake, NP with a determiner)

UNIVERSITÄT DES SAARLANDES



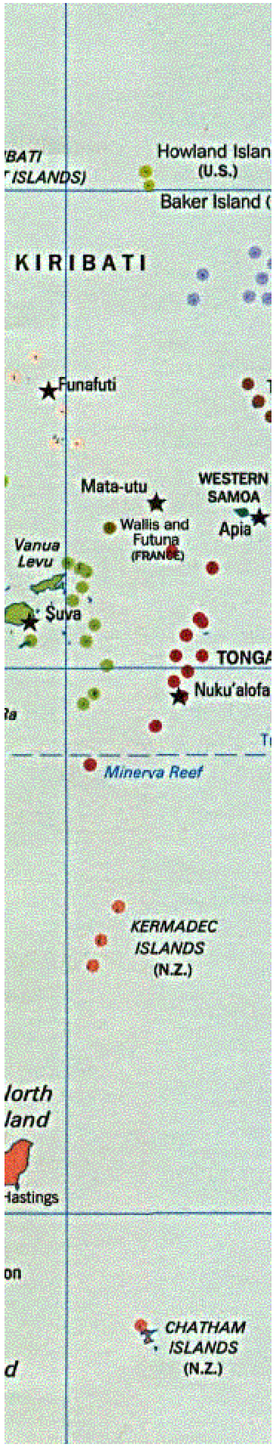
Sample Selection

They are really very different..

1. Pronoun
2. Definite Description
3. Proper Name

   name-matching, mainly NE-antecedents

4. Other (indefinite, bare plural, parsing mistake, NP with a determiner)



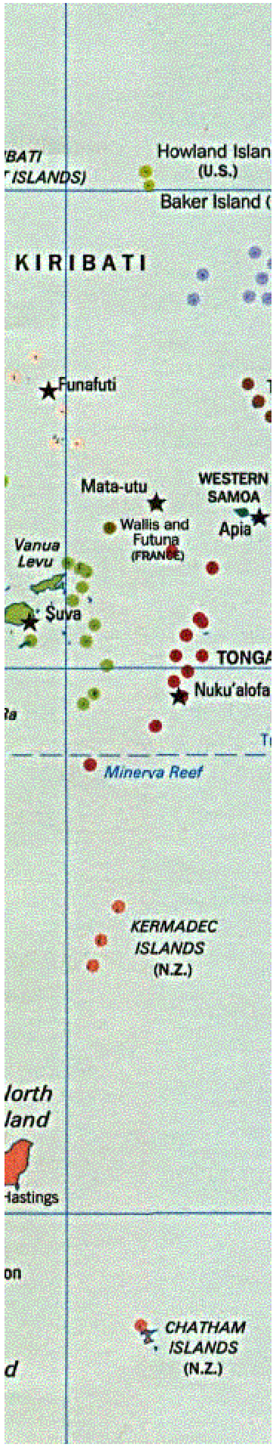
Sample Selection

They are really very different..

1. Pronoun
2. Definite Description
3. Proper Name
4. Other (indefinite, bare plural, parsing mistake, NP with a determiner)

    explicit indication for coreference, mainly discourse new

UNIVERSITÄT DES SAARLANDES



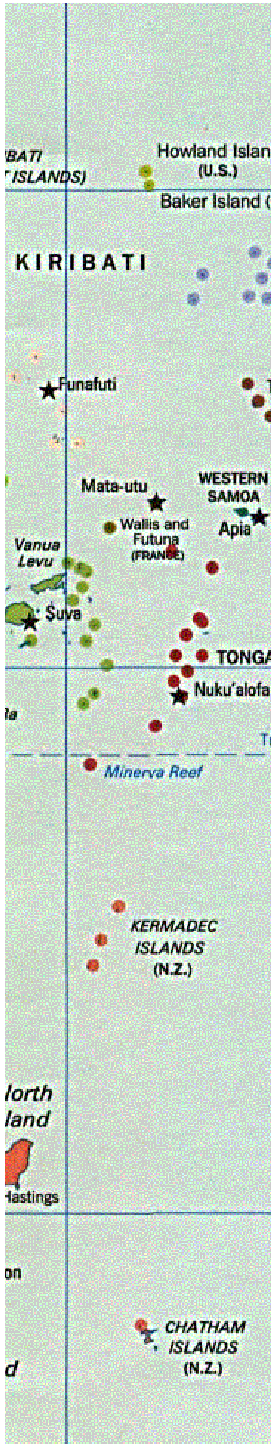
Sample Selection

Pronouns

Take all the close candidate antecedents.

Proximity criteria:

1. 2-sentence window
2. 5-sentence window
3. Same paragraph
4. <= distance (closest ante, anaph)



Sample Selection

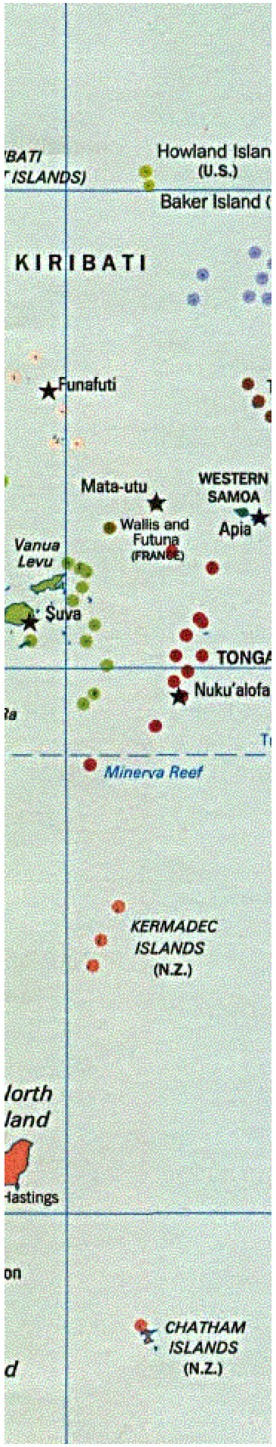
Definite descriptions

Look for a same-head candidate antecedent?

[+] -> include all the same-head antecedents + all the negatives between the closest one and the anaphor

[-] -> include all the non-pronominal positives; negative sample selection (Soon et al.)

UNIVERSITÄT DES SAARLANDES



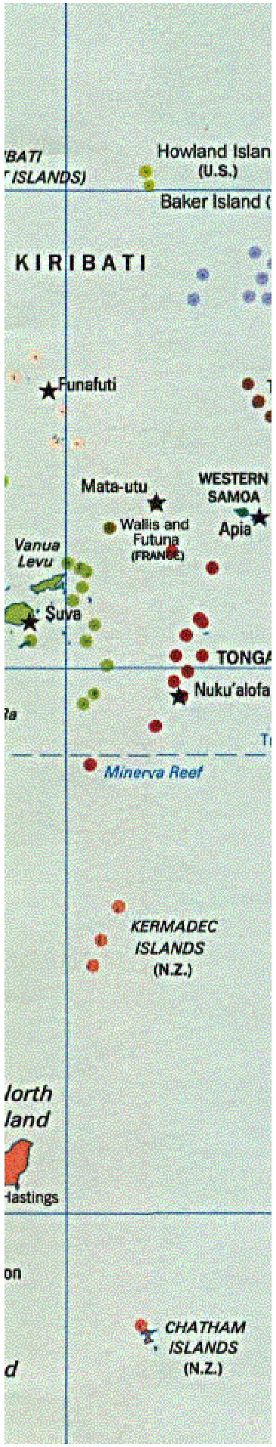
Sample Selection

Named Entities

Include only NE-antecedents

1. All
2. Apply Negative selection

UNIVERSITÄT DES SAARLANDES



Sample Selection

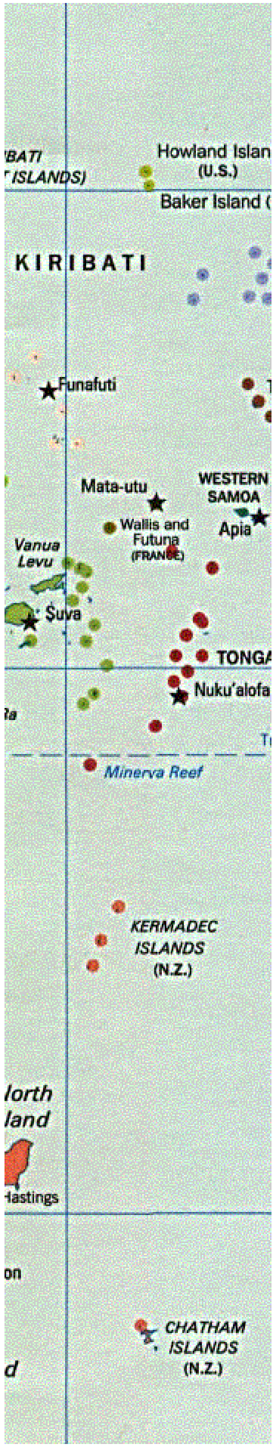
Remaining anaphors

Look for a construction, explicitly indicating coreference?

[+] -> include the antecedent + all the negatives between it and the anaphor

[-] -> discard

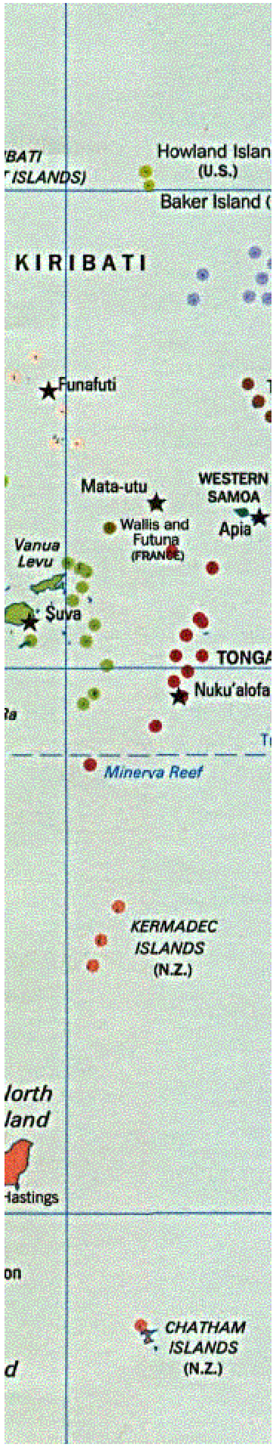
Explicit coreference constructions: appositions, copulas,..



Preliminary Results

	No Sample Selection	Sample Selection
Number of training instances	495144	147064
Learning time (CPU), sec	13435.11	4691.00
Recall, %	36.5	50.8
Precision, %	70.0	60.6
F-score, %	48.0	55.3

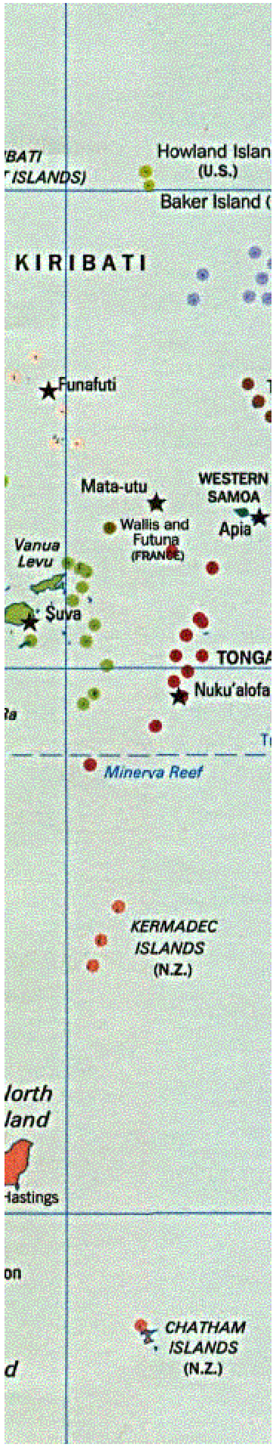
UNIVERSITÄT DES SAARLANDES



Conclusion

- Standard training data generation procedure is too simplistic: too many negative and too hard positive instances
- Different re-sampling for different types of anaphors
- Improves both the system's performance and speed

UNIVERSITÄT DES SAARLANDES



Future Work

- Feature Selection
- Clustering?

UNIVERSITÄT DES SAARLANDES