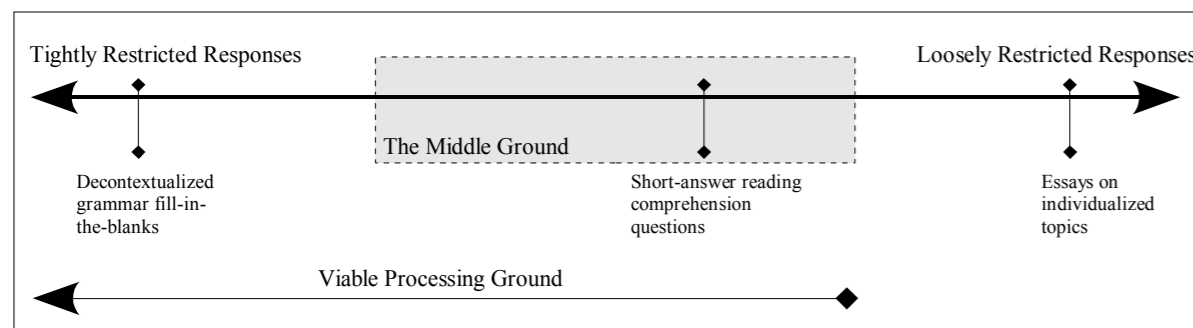


## Motivation

- Meaningful interaction in the foreign language is crucial for language learning.
- To be able to offer a wider range of activities, ICALL systems must be able to evaluate aspects of meaning.



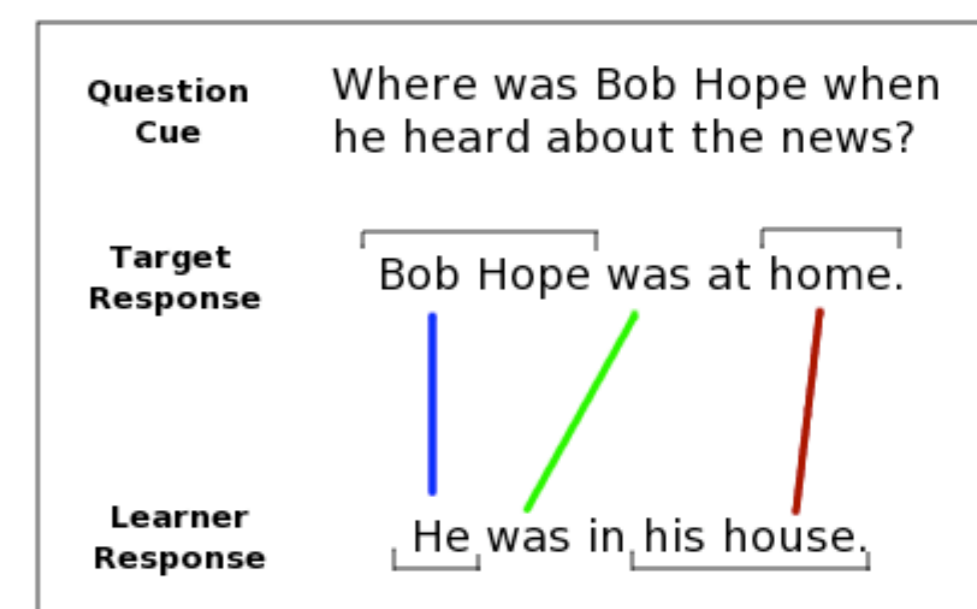
- Loosely restricted reading comprehension (RC) questions are a good test case:
  - Common activity in real-life foreign language teaching.
  - Responses can exhibit variation on lexical, morphological, syntactic, semantic levels.
  - It is possible to specify target answers.

## Data

- Learner corpus: 566 responses to RC questions from intermediate ESL students.
    - Development set: 311 responses from 11 students to 47 questions
    - Test set: 255 responses from 15 students to 28 questions
  - Graders provided target sentences, keywords.
  - Two graders annotated the data in two ways:
    - detection: Correct/Incorrect meaning
    - diagnosis (5 codes): correct; missing concept, extra concept, blend, non-answer
- Eliminated 31 responses (12%) which the graders did not agree on.
- On average, 2.7 form errors per sentence.
  - Learner responses vary significantly; no full string or bag-of-word overlap with targets.

## Method

- Comparison of target and learner responses on token, chunk and relation levels, e.g.,:



- 14 Features:
  - keyword/head overlap
  - target token overlap
  - learner token overlap
  - target chunk overlap
  - learner chunk overlap
  - target triple overlap
  - learner triple overlap
  - % token matches
  - % lemma matches
  - % synonym matches
  - % similarity matches
  - % sem. type matches
  - match variety
  - (sem. error detection)
- Combination of features computed by
  - manual rules
  - machine learning (TiMBL), using majority voting on available distance measures

### Loosely restricted reading comprehension questions: An example

**Question:** What are the methods of propaganda mentioned in the article?

**Target:** The methods include use of labels, visual images, and beautiful or famous people promoting the idea or product. Also used is linking the product to concepts that are admired or desired and to create the impression that everyone supports the product or idea.

**Learners:**

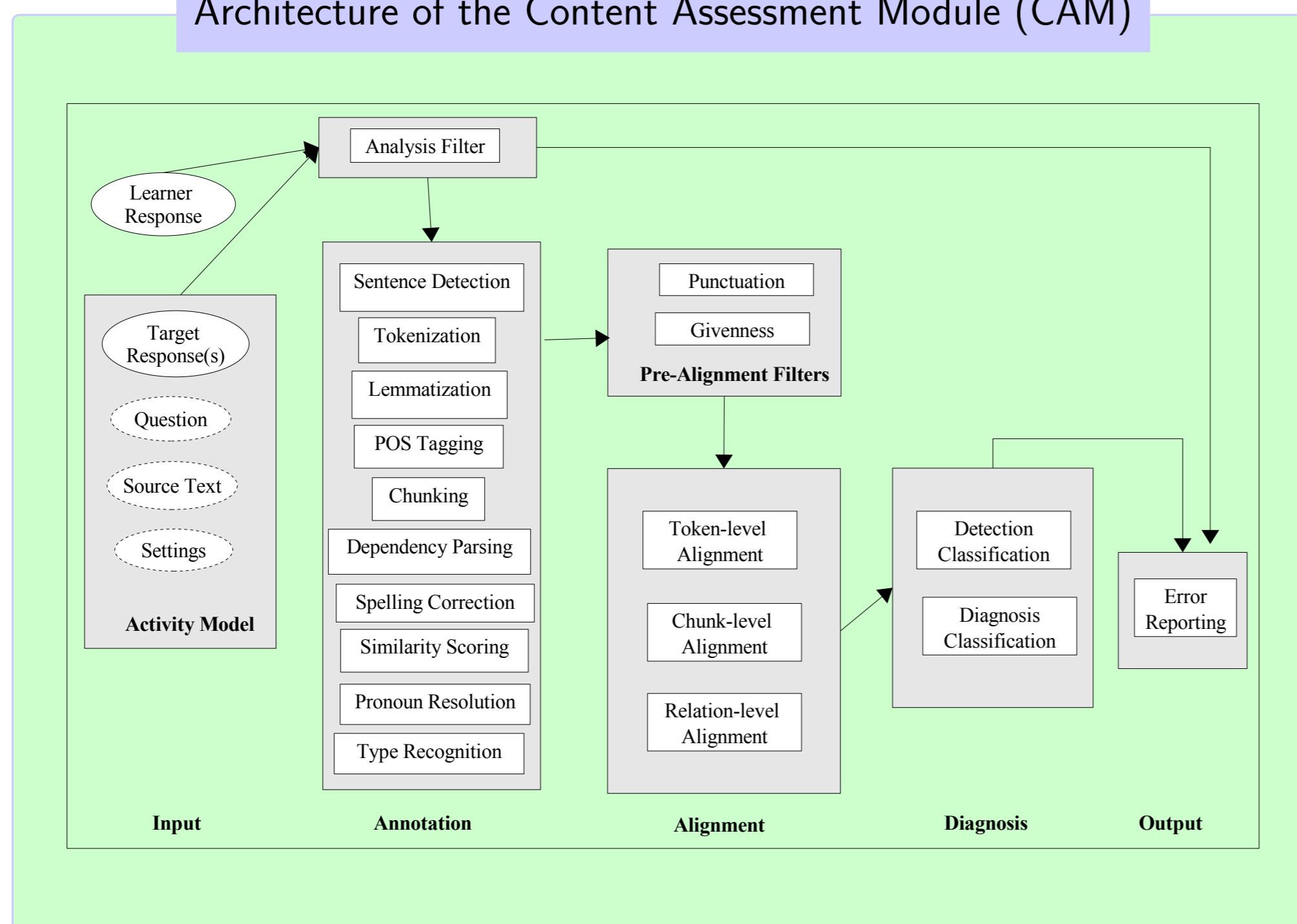
- A number of methods of propaganda are used in the media.
- Bositive or negative labels.
- Giving positive or negative labels. Using visual images. Having a beautiful or famous person to promote. Creating the impression that everyone supports the product or idea.

## Results

Detection	Accuracy
Baseline	50%
Development Set: Manual CAM	81%
Development Set: CAM	87%
Test Set: Manual CAM	63%
Test Set: CAM	88%

Diagnosis with 5 codes	Accuracy
Development Set	87%
Test Set	87%

### Architecture of the Content Assessment Module (CAM)



- The good performance confirms the viability of using shallow NLP techniques for meaning error detection.
- Form errors don't negatively impact results:
  - 68% of correctly diagnosed had form error
  - 53% of incorrectly diagnosed ones did so.
- Even for small data sets, machine learning can benefit shallow content assessment.
- No directly comparable systems exist, but the results are competitive given, e.g., 85% accuracy obtained by C-rater (Leacock 2004), an automatic scoring system for short answers written by native speakers.