A Cross-Language Question/Answering-System for German and English

Günter Neumann & Bogdan Sacaleanu
Language Technology Lab
German Research Center for Artificial Intelligence
DFKI, Saarbrücken

Major motivation (*the big pic.*)

The idea is that depending on the complexity of the query information (from simple fact-based questions, to relational template-based questions, to thematic-oriented questions) shallow or deep QA strategies should be selected (or even mixed) which might involve different degrees of linguistic processing, domain reasoning or interactivity between a user and the system.

Large-scale hyprid QA technology
(BMBF-funded project Quetal, starting March 2003)

1. Open-domain/domain-specific
2. Cross-language
3. Semi-structured documents

Research goal: Adaptive QA-technology

1. Linguistic core technology (variable-depth „text-zooming")
2. From shallow to deep
3. Large-scale Machine Learning (ontology bootstrapping, anticipation feedback loops)

Starting point for our QAatCLEF: system BiQue

- Foster bottom-up system development
  - Data-driven
  - Robustness
  - Scalability
  - Uniform NLP components
- Shallow answer processing
  - Bag of objects
  - Sentence-based equivalence classes
- Initial work:
  - Monolingual German Web-based answer extraction system
  - Evaluation against Google snippets (Neumann&Xu, 2003)

Common LT-core for bilingual query & answering processing

Bag of Object approach (cf. Light et al. 2002)

- Shallow semantic representation
- Set of linguistic objects (partially indexed)
- Set operations for overlap & equivalence class

NE-recognition (unsupervised ML based on CollinsSinger, 1999)

- Generic linguistic principles for NE-candidate identification (president of XXX -> XXX=company)
- Subclassification of ShProT's phrase types
- Based on learned decision list

Tree-based grammar for clause level

- Lexicalized tree grammars
- Subgrammars for query analysis
- HPSG-DOP as linguistic basis (Neumann, 2003; not yet integrated)

Shallow processing with ShProT (Brants&Skut)

- Statistical-based PoS & phrase tagging
- XML-API

Bag of Objects

B := {O_1,…,O_n;α}, where: $O_i \neq O_j$ for $i \neq j$ & weight α

Objects:

NE:
PoS:
Stem:
Wf:
Weight:

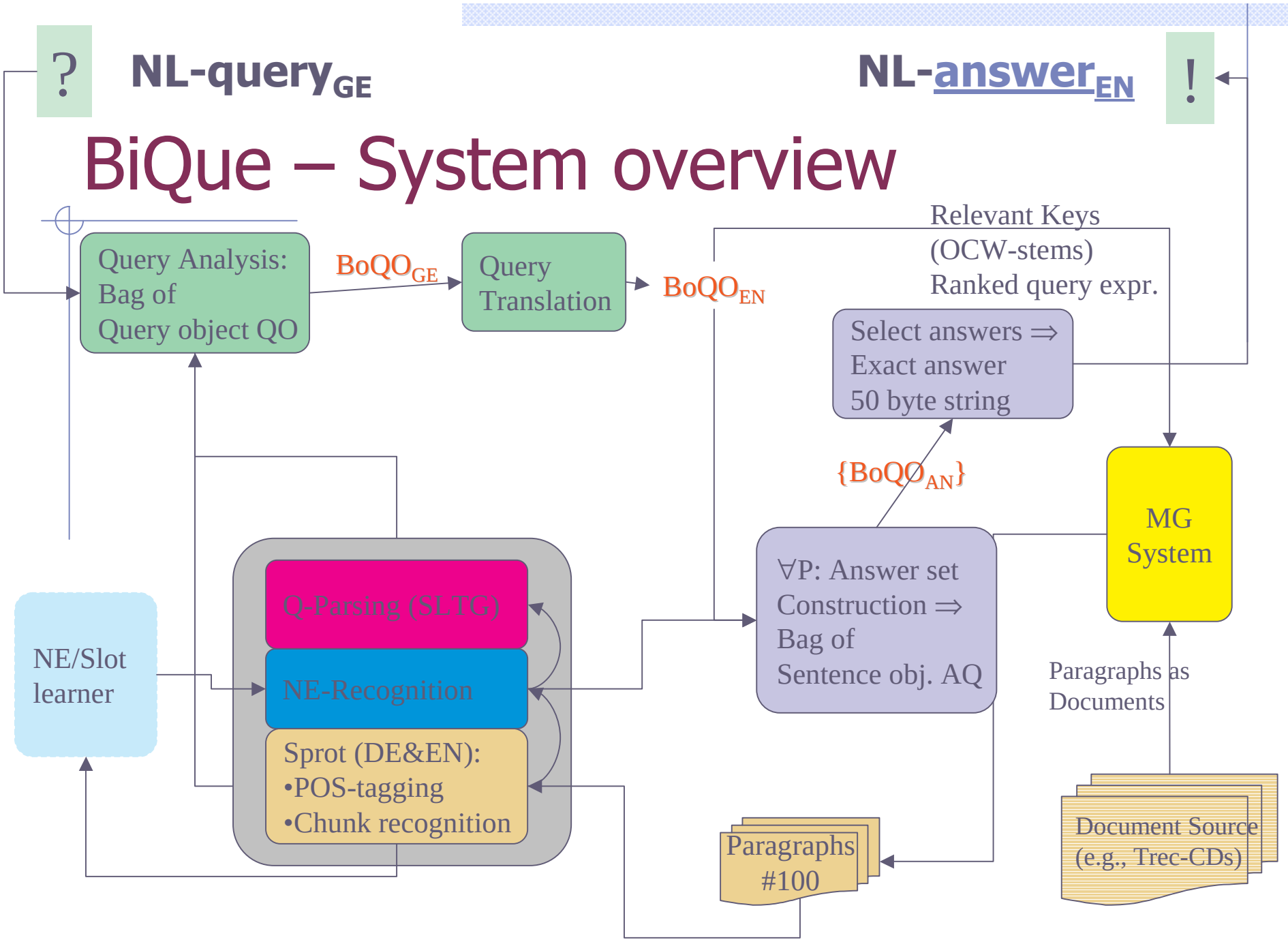
Objects are word-based entities;
the structure can also be seen as a generic API; weights are given a priori.
Objects can also be linked.

Overlap $Ov_{s1,s2} = B_{s1} \cap B_{s2}$

Two Objects O_i,O_j match if:
same lemma & PoS; $\alpha=\alpha(O_i)$)
Or same lemma; $\alpha=\alpha\ (O_i)/2$
Or same wordform; $\alpha=\alpha\ (O_i)/2$;
$\alpha(Ov)=\Sigma\alpha(O_i)$
NE is taken into account later

Overlap set of a query q

$Os_q = \{s_1,\ldots,s_n\}$, with: $Ov_{q,s_i} = Ov_{q,s_j}$ for $i \neq j$

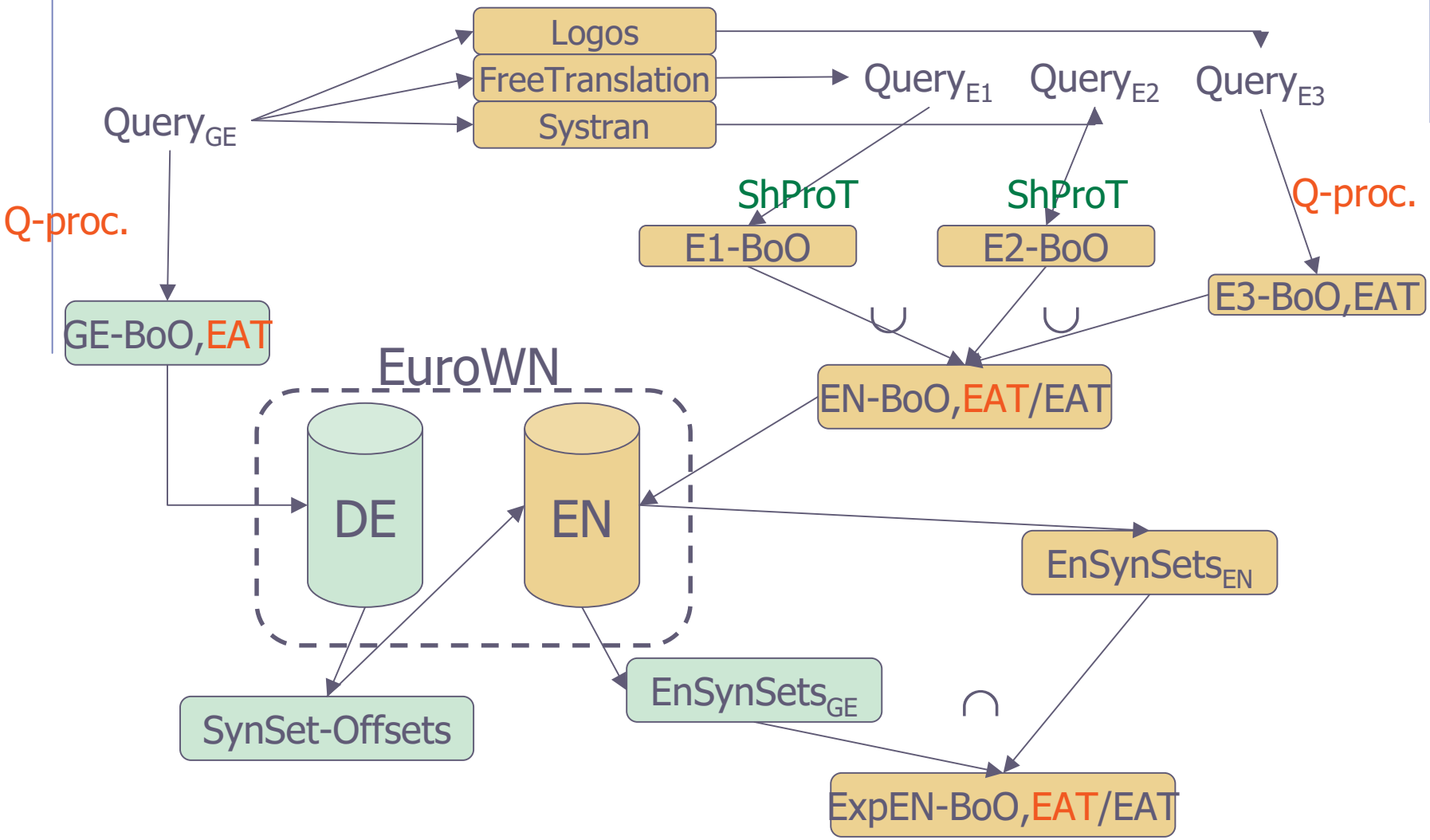

?
NL-query$_{GE}$
NL-answer$_{EN}$
!
BiQue – System overview
Query Analysis:
Bag of
Query object QO
BoQO$_{GE}$
Query
Translation
BoQO$_{EN}$
Relevant Keys
(OCW-stems)
Ranked query expr.
Select answers ⇒
Exact answer
50 byte string
{BoQO$_{AN}$}
MG
System
∀P: Answer set
Construction ⇒
Bag of
Sentence obj. AQ
Paragraphs as
Documents
NE/Slot
learner
Q-Parsing (SLTG)
NE-Recognition
Sprot (DE&EN):
•POS-tagging
•Chunk recognition
Paragraphs
#100
Document Source
(e.g., Trec-CDs)

Query translation & expansion

- First idea:
  - Only use EuroWordNet
  - Defines a word-based translation via synset offsets
- Experience
  - EuroWordNet too sparse on German side
  - Neverless introduced too much ambiguity
  - NE-translation is crucial
- So far, not very much of help

- New idea:
  - Use EuroWordNet
  - Use external MT-services
  - Overlap-mechanism for query expansion
- Crosslingual because
  - EAT from GE-QueryBoO
  - Synsets from EuroWN direct query expansion
- Experience
  - External MT services used ako WSD
  - Reduced degree of ambiguity

Query translation & expansion



Example

1. Translation services as WSD

Wo wurde das Militärflugzeug Strike Eagles 1990 *eingesetzt*?

FreeTranslation: *Where did the military airplane become would strike used Eagles 1990?*

Systran: *Where was the military aircraft Strike Eagle used 1990?*

Logos: *Where was the soldier airplane Strike Eagles installed in 1990?*

BoO_{EN} := {soldier, airplane, strike, eagle, install, 1990, military, become, strike, *use*, aircraft}

2. Query expansion using EuroWN

$\forall x \in BoO_{EN}$: lookup(EuroWN);
If x is unambiguous: extend BoO_{EN}
Else $\forall$readings(x):
get its alligned German readings &
Look them up in BoO_{GN}
If successfully then add English terms to BoO_{EN}

~~Reading-697925~~
~~EN: {handle, *use*, wield}~~
~~DE: {handhaben, hantieren}~~

~~Reading-1453934:~~
~~EN: {behave toward, use}~~
~~DE: not aligned~~

Reading-658243:
EN: {apply, employ, make use of, put to *use*, use, utilise, utilize}
DE: {anbringen, anwenden, bedienen, benutzen, *einsetzen*, …}

Results

- Very first participation at Trec-like event
- Focus on system implementation rather than system adaptation
- Submitted only one 50byte run
  - 14.5 % (strict), 15% (lenient)
  - No processing of questions with quoted terms
- Tried second run:
  - Preprocessing of whole corpus with ShProT in order to bypass MG's stemmer
    - (e.g., Belize -> Bel & Bell -> Bel)
  - Too time consuming
    - Online, incremental approaches needed