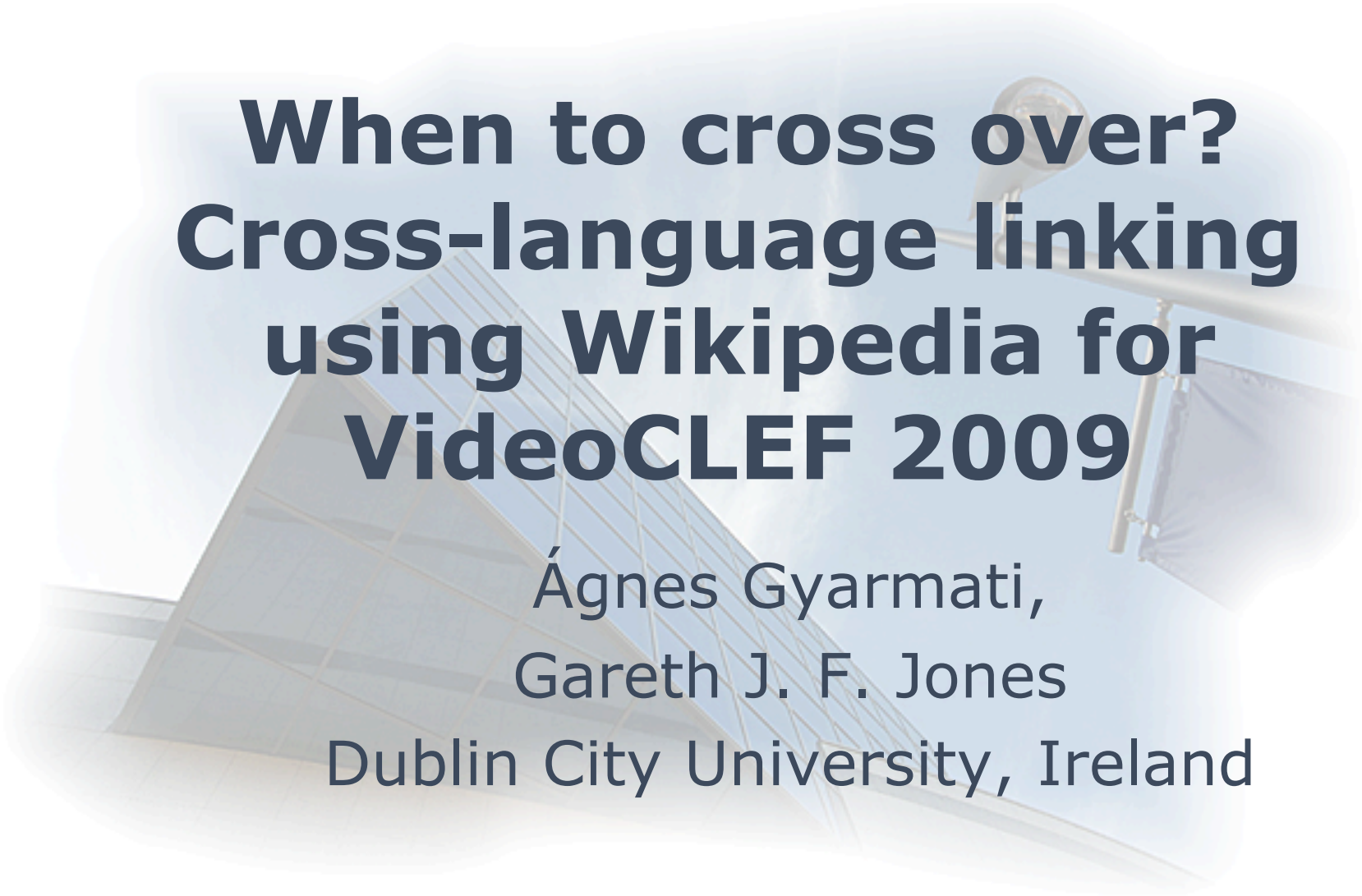

When to cross over? Cross-language linking using Wikipedia for VideoCLEF 2009

Ágnes Gyarmati,
Gareth J. F. Jones
Dublin City University, Ireland

Outline

- The task
- Our solution
  - in theory (the idea)
  - in practice (the implementation)
- Results
- Problems
- Further work

The task

- VideoCLEF 2009: "Linking Task"
  - search for pages in English Wikipedia related to Dutch language documentary videos
- complex task – complicated solution
  - viewed as a sequence of subtasks
  - 2 basic variants
    - the difference: language switch on a different point of the workflow

How - in theory (1)

- Extract query from the transcript (NL)
- Use query for a search in Wikipedia (NL)
- Check Wikipe → d ia's own cross-language links (NL EN)
- Return list of relevant pages (EN)

How - in theory (2)

- Extract query from the transcript (NL)
- Translate query (NL → EN)
- Use the
  tran
  slated query to search in Wikipedia (EN)
- Return list of relevant pages (EN)

How - in practice: Collections

- Wikipedia dump (30/31. May 2009)
  - NL & EN
  - basic cleanup
- Indexing (and retrieval) with Lemur
  - Lemur's built-in language model Indri
  - stemming:
    - EN stemmed (built-in option in Lemur)
    - NL unstemmed
    - NL stemmed: Porter stemmer for Dutch
      - algorithm: Snowball
      - implementation: Oleander
  - stopping: Snowball's stopword lists

How – in practice: Queries

- Extract query from the transcript (NL):
  - a simple sequence of words spoken
    b
    etween starting & end point of anchor points
  - stemmed - if collection is stemmed
- Query translation - if needed
  - Multimatch
    - WorldLingo machine
      tr
      anslation + cultural heritage vocabulary extension

How – in practice: Retrieval

- Indri

R ↰ •

  - none
  - blind relevance feedback
    - Indri's default model
      - 10 documents, 5 expansion terms

Results

- For primary links

Run	*Recall*	*MRR*
NL Wiki	**44/165**	**0.182**
NL Wiki – stemmed	**44/165**	**0.182**
EN Wiki – stemmed	13/165	0.056
NL Wiki + relevance feedback	38/165	0.144

Results

- For related links
  - primary & secondary links
  - no Recall defined

Run	*MRR*
NL Wiki	0.268
NL Wiki – stemmed	**0.275**
EN Wiki – stemmed	0.090
NL Wiki + relevance feedback	0.190

Problems (1)

- search performed in Dutch Wikipedia
  - generally:
    - ASR
  - specifically:
    - missing links: no equivalent page in English Wikipedia

Problems (2)

- search performed in English Wikipedia
  - generally:
    - ASR
  - specifically:
    - machine translation
    - size of collection
    - irrelevant data left in collection

Further work

- more thorough cleanup
  - one further step done in EN:
    - primary links: 40/165 (inofficial preliminary result)
- other methods in query formation
- other models in relevance feedback

Thank you