CILC 2013
International Conference on Corpus Linguistics
V Congreso Internacional de Lingüística de Corpus

AUTOMATIC ACCESS TO LEGAL TERMINOLOGY APPLYING TWO DIFFERENT ATR METHODS

MARÍA JOSÉ MARÍN
CAMINO REA

UNIVERSIDAD DE MURCIA

1.INTRODUCTION

-Importance of identifying the terms in a specialised corpus.

- Terms are "textual realisations of specialised concepts" (Spasic et al. 2005).

-They are employed to communicate amongst specialists (Rea, 2008).

- They are mono-referential and have a univocal character (Cabré, 1993): form ⟹ content.

-Potential applications of automatic term recognition (ATR): building dictionaries and glossaries; machine translation; ontology building, etc.

1.INTRODUCTION

- This paper evaluates the efficiency of two ATR methods: *Keywords* (2008) and Chung's (2003) on a 2.6m word legal corpus *(UKSCC).*

- Both will be validated in terms of precision and recall.

2. *UKSCC* and *LACELL:* the study and reference corpora

- *UKSCC (United Kingdom Supreme Court Corpus):*

  - Legal corpus of 192 judicial decisions issued by the Supreme Court of the United Kingdom (2008-2010).

  - Compiled *ad hoc* according to CL standards (Sánchez, 1995; Wynne, 2005; Pearson, 1998; Rea, 2010).

  - Mononolingual and synchronic.

  - The Supreme Court chosen as source of texts due to its importance as a judicial institution: touches upon all branches of law and greater geographical scope.

  - Judicial decisions appear as the main source of law in common law countries.

2. *UKSCC* and *LACELL:* the study and reference corpora

- ***LACELL (Lingüística Aplicada Computacional, Enseñanza de Lenguas y Lexicografía) :***

  - **Balanced general English corpus of 20m words: written (newspapers, books, magazines, brochures, letters, etc.) and oral language samples.**

  - **Compiled by the LACELL research group at Murcia University (English Dept.).**

3. ATR method description

Automatic term recognition (ATR) methods date back to the 1980s: they allow handling large amounts of data automatically spotting the most relevant terms in a specialised corpus. They have been profusely reviewed (Maynard and Ananiadou, 2000; Cabré et al., 2001; Drouin, 2003; Lemay et al., 2005; Vivaldi et al., 2012, etc.)

- *Keywords* (Scott, 2008):

  - Not an ATR method proper, however, it has proved to identify legal terms more efficiently than other methods designed to that purpose.

  - Automatically implemented using *Wordsmith 5.0.* Settings adjusted tu use Dunning's (1993) log-likelihood algorithm.

3. ATR method description

- Chung's method (2003)

  - *S*ingled out due to high rate of success recorded by the author (86% precision on average).

  - Chung compares a qualitative term recognition method: *the rating scale approach* with her own quantitative one. She concludes that terms displaying a > 50 ratio of occurence are terms.

  - How to calculate it:

  Wr= SF(w)/RF(w) (freq. counts must be normalised)

4. Validation process and results

4.1. Validation process

-Precision: Percentage of true terms out of candidate terms extracted.

-Recall: Percentage of true terms out of total amount of terms in the whole corpus.

-We resorted to automatic validation: 10,000 entry legal electronic glossary compiled by authors used as gold standard (human validation poses problems due to subjectivity) .

-Terms confirmed as true if found in glossary.

-*Keywords* implemented automatically; Chung's method applied using spreadsheet (data obtained with *Wordsmith* too)

4. Validation process and results

4.2. Results

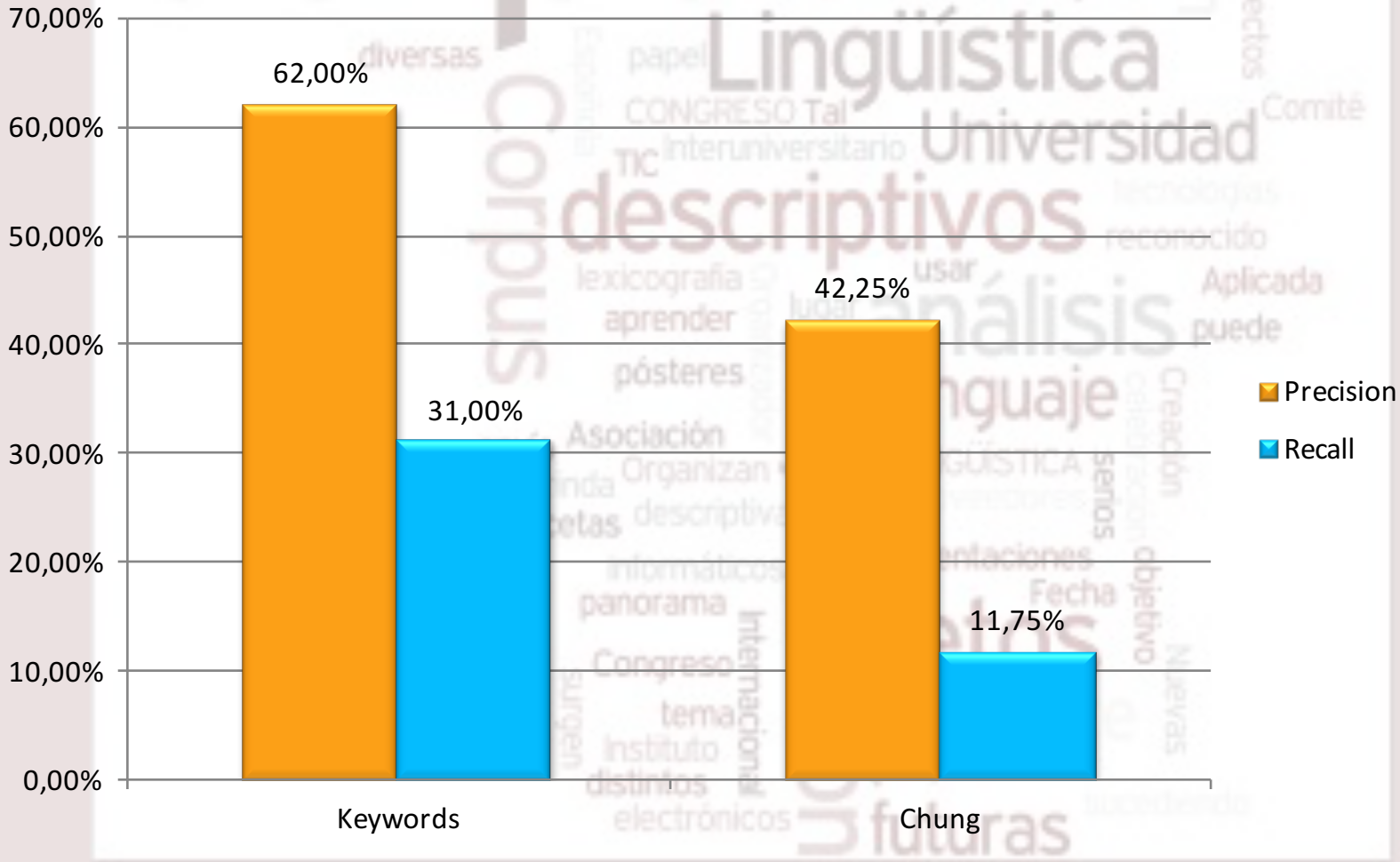

Fig. 1 *Overall precision and recall achieved by each method*

4. Validation process and results

4.2. Results

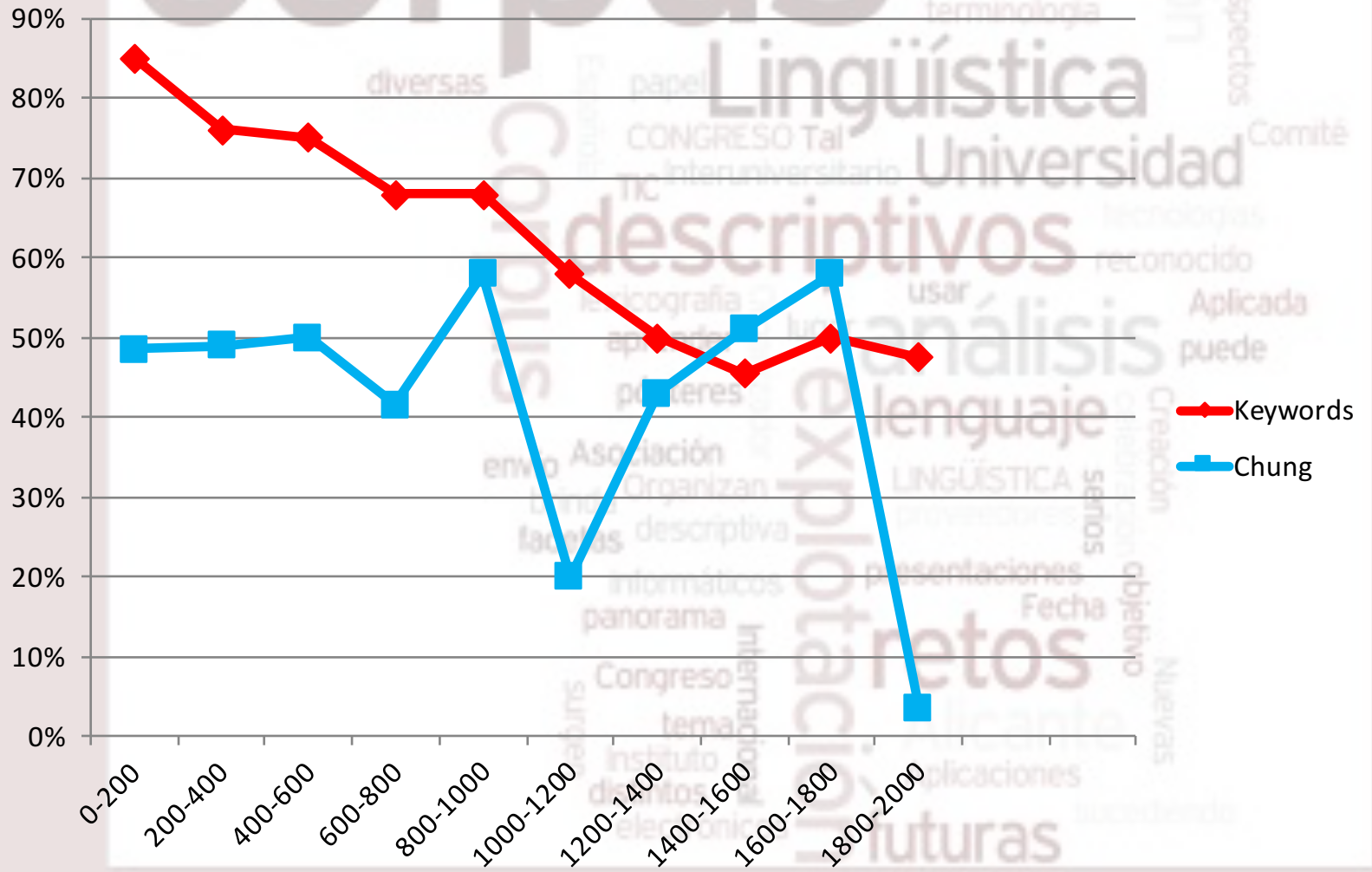

Fig.2 *Cumulative precision attained on top 2000 candidates*

4. Validation process and results

4.2. Results

-*Keywords* excels Chung's method both in terms of precision and recall.

-*Keywords* decreases its efficiency smoothly and constantly whereas Chung's method performance is much more irregular.

-Chung's method only performs better within candidates 1600-1800 dropping sharply to 3.5% precision afterwards.

-Chung's bad results may be due to the automatic inclusion of words not in the reference corpus in the terms group, especially proper names so typical of judicial decisions.

4. Validation process and results

4.2. Results

Table 1. *Top 25 candidate terms extracted by each method*

Chung's method (2003)	Ratio	Keywords (2008)	Keyness
EHRR	∞	**COURT**	28955.793
EWCA	∞	SECTION	27627.5586
UKHL	∞	**PARA (paragraph)**	25311.1152
MANCE	∞	LORD	25155.4434
SIAC	∞	**V (versus)**	22486.0918
ECHR	∞	**APPEAL**	21236.8652
EWHC	∞	**ARTICLE**	19301.6328
BAILII	∞	**ACT**	18577.8652
GESTINGTHORPE	∞	**CASE**	18328.9512
FOSCOTE	∞	**LAW**	10458.0918
EARLSFERRY	∞	**JUDGMENT**	9297.75
JFS	∞	**APPELLANT**	8048.33496
ECTHR	∞	**PROCEEDINGS**	7787.61963
STOJEVIC	∞	**CONVENTION**	7764.64355
TURPI	∞	WHETHER	7716.16992
LJ	∞	**LJ**	7707.0918
DALLAH	∞	**RIGHTS**	7023.53613
SUMPTION	∞	**DECISION**	6950.50488
SEISED	∞	**ORDER**	6632.18164
BANKOVIC	∞	**JURISDICTION**	6374.33105

5. Conclusion

- Evaluating the efficiency of ATR methods is highly recommendable to select the ones that suit our corpus best, especially due to the fact that some of them are domain-dependent like Chung's .

-Nevertheless, as put forward by Lemay (2005: 245), "much still remains on the terminologist's ability to differentiate a relevant unit from a non-relevant one. Lists must be scanned to remove irrelevant units". Actually, "fine-grained semantic distinctions still rely ... on terminologists".

References

Alcaraz Varó, E. (2000) *El inglés profesional y académico*. Madrid: Alianza Editorial.

Cabré, M.T. (1993). *La terminología. Teoría, metodología, aplicaciones.* Barcelona: Antártida/ Empúries.

Cabré, M. T., Estopà, R., Vivaldi, J. (2001). Automatic term detection: a review of current systems. Bourigault, D., Jacquemin, C., L'Homme, M.C. (Eds.). *Recent Advances in Computational Terminology* 2, 53-87. Amsterdam: John Benjamins, Natural Language Processing.

Chung, T. M. (2003). A corpus comparison approach for terminology extraction. *Terminology* 9(2), 221-246.

Heatley, A., Nation, I.S.P. 1996. *Range* (computer software). Wellington: Victoria University of Wellington.

Kit, C. and Liu, X. (2008) "Measuring mono-word termhood by rank difference via corpus comparison". *Terminology*, 14 (2): 204-229.

Lemay, C., LHomme, M.C., Drouin, P. (2005). Two methods for extracting "specific" single-word terms from specialized corpora: experimentation and evaluation. *International Journal of Corpus Linguistics,* 10(2), 227-255.

Marín, M.J., Rea, C. (2011). "Design and compilation of a legal English corpus based on UK law reports: the process of making decisions". Carrió Pastor, M. L., Candel Mora, M.A. (Eds.). *Las Tecnologías de la Información y las Comunicaciones: Presente y Futuro en el Análisis de Córpora*. *Actas del III Congreso Internacional de Lingüística de Corpus* (101-110). Valencia: UniversitatPolitècnica de València.

Maynard, D. and Ananiadou, S. 2000. TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing* 8(1), 101–125.

Pearson, J. (1998). *Terms in Context*. Amsterdam: John Benjamins Publishing Company.

Rea, C. (2008). *El Inglés de las Telecomunicaciones: Estudio Léxico Basado en un Corpus Específico*. Tesis doctoral. Murcia: Universidad de Murcia.

Rondeau, G. (1983). *Introduction à la terminologie.* Québec: Gaëtan Morin Editeur.

Sánchez, A., Cantos, P., Sarmiento R., Simón, J. 1995 *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

Scott, M. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Spasic, I., Ananiadou, S., McNaught, J. & Kumar, A. 2005. 'Text mining and ontologies in biomedicine: Making sense of raw text'. *Brief Bioinform*, 6(3), 239-251.

Vivaldi, J., Cabrera-Diego, L.A., Sierra, G., Pozzi, M. (2012). 'Using Wikipedia to Validate the Terminology Found in a Corpus of Basic Textbooks'. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Instambul, May 2012.

Wynne, M. (Eds.) 2005 *Developing Linguistic Corpora: a Guide to Good Practice.* ASDS Literature, Languages and Linguistics. Oxford.