UNIVERSITAT JAUME·I

Jornades de Foment de la Investigació

ANÁLISIS DE LA EXTRACCIÓN Y VALIDACIÓN BILINGÜE DE TERMINOLOGÍA CON EL PROGRAMA INFORMÁTICO MULTITERM EXTRACT

Autors

Manuel RUBIO
Verónica PASTOR
Esperanza VALERO.

INTRODUCCIÓN

Este trabajo se enmarca dentro del proyecto CREC V: *Creación de recursos lingüísticos electrónicos y adquisición de destrezas en tecnologías de la terminología y la traducción*[1]. Este proyecto tiene como objetivos la creación de recursos lingüísticos electrónicos y la adquisición por parte del estudiantado de destrezas en tecnologías de la terminología y la traducción. En ediciones anteriores de este proyecto, una de las tareas fue la creación de un corpus bilingüe alineado (inglés-español) del ámbito de la cerámica. A partir de este recurso se puso en marcha una nueva tarea todavía en curso para extraer y validar terminología bilingüe con la herramienta Multiterm 2007 Extract.

En el marco del proyecto CREC se efectuó un estudio anterior (Benavent y Parrilla, 2007) en el que se analizó la eficacia de la herramienta ExtraTerm en la extracción monolingüe de terminología cerámica. En este artículo presentamos un estudio piloto en el que analizamos los problemas que genera el extractor Multiterm 2007 Extract en la extracción de terminología bilingüe.

EXTRACCIÓN BILINGÜE DE TERMINOLOGÍA

La extracción automática de terminología es una importante área de investigación debido a su aplicación en la creación de recursos como diccionarios o glosarios especializados, sistemas de procesamiento del lenguaje natural, traducción automática, etc. Por ello, la incorporación de los programas de extracción automática en el programa de la asignatura de Terminología resulta de gran utilidad para el futuro profesional de los traductores (Alcina, 2003).

Las herramientas de alineación y de extracción de concordancias bilingües están en un nivel avanzado de desarrollo y permiten obtener resultados de calidad sin demasiada intervención manual. Sin embargo, la extracción automática de léxico bilingüe todavía no ha alcanzado el mismo nivel (Kraif, 2008: 84). Además, aunque actualmente se puede encontrar gran cantidad de textos multilingües en diversos ámbitos del conocimiento con los que se podría extraer terminología bilingüe para la creación de diccionarios especializados, la realidad es que los diccionarios disponibles todavía son escasos (Déleger, 2006: 9).

Son numerosos los autores que estudian los problemas que generan las herramientas de extracción automática de terminología monolingüe (Estopà, 1999, Vivaldi y Rodríguez, 2007) y bilingüe (McEnery et al., 1997, Névéol y Ozdowska, 2005,). Como señalan estos autores, el ruido y el silencio son

1 Este proyecto se ha desarrollado durante el curso académico 2008-2009. Es un proyecto financiado por la Unitat de Suport Educatiu (USE) de la Universitat Jaume I. La directora del proyecto es Amparo Alcina.

dos problemas que existen en todos los extractores de terminología. El ruido ocurre cuando el extractor propone candidatos a término o equivalencia que no son válidos en un determinado campo de especialidad. Por el contrario, el silencio consiste en la omisión en la extracción de algunos términos propios del campo de especialidad. El estudio de estos dos fenómenos es importante para la mejora de las herramientas de extracción automática.

RECURSOS UTILIZADOS

Corpus electrónico alineado (inglés-español) de la cerámica

Este corpus procede de la digitalización y posterior alineación de las actas de los distintos congresos mundiales de la calidad del azulejo y pavimento cerámico Qualicer. Este corpus está integrado por textos en formato txt que contienen los segmentos alineados en inglés y en español etiquetados con el formato de Trados Workbench (Soler, Alcina y Estellés, 2006; Alcina et al., 2007).

Herramienta de extracción automática Multiterm Extract

El recurso que utilizamos para la extracción de los candidatos a término fue la herramienta de extracción automática Multiterm 2007 Extract, de SDL. Este programa se basa principalmente en un método estadístico, que extrae la terminología en función de su frecuencia de aparición en el corpus y aplica también algunas reglas lingüísticas. Además, Multiterm 2007 Extract incluye un archivo de exclusión o stop-list que sirve para excluir de la extracción algunas palabras del lenguaje general, que aunque aparecen con frecuencia en el corpus, no son términos. El programa presenta los candidatos a término y sus equivalentes en una lista de dos columnas con los candidatos a la izquierda y los equivalentes a la derecha. Una vez realizada la validación de terminología, el programa permite exportar los resultados a una base de datos terminológica en formato xml.

METODOLOGÍA

La metodología seguida en este trabajo se puede dividir en cuatro fases. Las tres primeras destinadas a la extracción y validación de terminología y la última al análisis de resultados. Estas fases son las siguientes:

- Extracción automática con Multiterm 2007 Extract.
- Validación manual de los candidatos a término y a equivalente.
- Revisión de la validación.
- Análisis de los datos y evaluación de resultados.

Extracción automática con Multiterm 2007 Extract.

Los estudiantes que colaboran en el proyecto CREC se encargan de la extracción automática de los candidatos a término y los equivalentes siguiendo los pasos que se especifican en el protocolo de extracción y validación bilingüe de terminología. Cada estudiante recibe un paquete de alineación con un archivo de texto que contiene los segmentos alineados en español e inglés. Crean un proyecto de extracción al que asignan un nombre del tipo VBce01-1, que indica que se trata de un proyecto de **v**alidación **b**ilingüe de la **ce**rámica. La lengua origen del proyecto es el español y la lengua meta el inglés. Las opciones seleccionadas para la extracción son las siguientes:

- La extensión mínima de los términos es de una palabra
- La extensión máxima de los términos es de seis palabras
- Se desactiva la función de máximo número de términos extraídos
- La función de máximo número de equivalentes extraídos se mantiene en cinco
- La frecuencia mínima de traducción se mantiene en uno

Validación manual de los candidatos a término y a equivalente.

A partir del proyecto de extracción el programa genera una lista de dos columnas. En la columna de la izquierda aparecen los candidatos a término en español y en la columna de la derecha, los candidatos a equivalente para cada candidato a término. En esta fase, los estudiantes deben, en primer lugar, decidir si un candidato es término del campo de la cerámica y, en el caso de que lo sea validar el candidato y aquellos candidatos a equivalentes propuestos por el programa que sean equivalentes del término validado.

Para decidir si un candidato es término del ámbito de la cerámica se siguieron cuatro criterios de L'Homme (2005) que se encuentran explicados en el protocolo de la tarea:

1. **Entidad del ámbito de la cerámica**: la unidad extraída designa una entidad del ámbito de la cerámica (ej. cemento, baldosa, granulometría).

2. **Unidad predicativa que reenvía a una entidad del ámbito de la cerámica**: se considerarán válidas las unidades predicativas –verbos, nominalizaciones, adjetivos, etc. – cuyos agentes o pacientes ya han sido admitidos como términos en función del criterio a. Por ejemplo, podemos considerar que la unidad predicativa *cocer* es un término, ya que se combina con el paciente *bizcocho* que refiere a una entidad del ámbito de la cerámica. Podemos verlo en el ejemplo: "el bizcocho se cuece en un horno de cocción rápida". Sin embargo, si al combinar *producción industrial* con *muebles* mantiene el mismo sentido que al combinarlo con *baldosas cerámicas* no se considerará válido.

3. **Derivado morfológico de un término**: si se trata de derivados morfológicos se considerarán válidos si están semánticamente emparentados con un término seleccionado en función de los criterios A o B (ej. *sellar: sellado, sellante, sellador; cocción: cocer, bicocción, cocido*).

4. **Unidad que establece una relación paradigmática con otro término**: si se trata de una unidad léxica que tiene una relación paradigmática (que no sea una relación morfológica ya identificada en C) con un término seleccionado en función de los criterios A, B o C, se considerará válido. Por ejemplo, si el adjetivo *extrusionado* se considera válido porque se combina con agentes especializados (aplicación del criterio B), *extrudido* deberá igualmente considerarse válido porque se consideran sinónimos (ej. *gres extrusionado; gres extrudido*).

Para aplicar estos criterios los estudiantes se ayudan de los contextos del corpus, a los que se puede acceder desde el botón *Concordance* del programa Multiterm Extract.

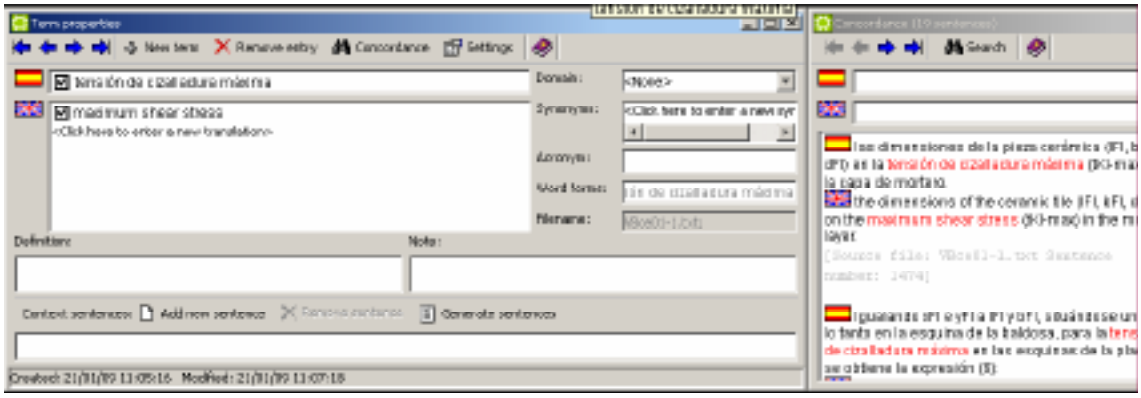

Ilustración 1. Opción Concordance del programa Multiterm Extract.

Si los estudiantes deciden que un candidato es término, lo validan activando la casilla que aparece a la izquierda de dicho término, como podemos ver en la ilustración 2. Asimismo, el término aceptado debe aparecer en su forma canónica, es decir, sin morfemas flexivos ni desinencias verbales. Por último, el programa propone una lista de candidatos a equivalentes para cada candidato a término. Los estudiantes también seleccionan las equivalencias de los términos validados. Al igual que el término en lengua original, las equivalencias también deben aparecer en su forma canónica.

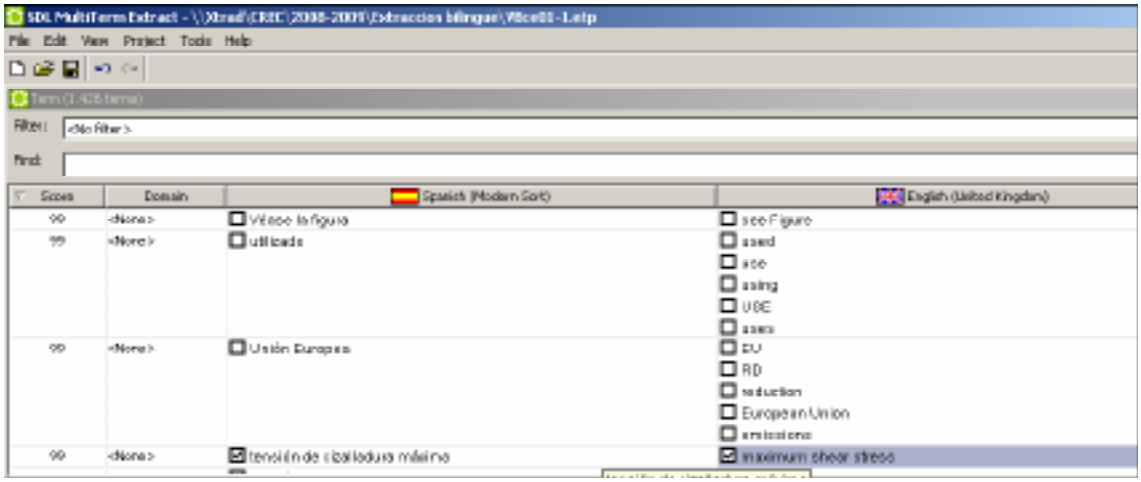

Ilustración 2. Ventana para la validación de términos y equivalentes del programa Multiterm Extract de SDL.

Revisión de la validación.

Una vez terminado el proceso de validación de los estudiantes, el personal colaborador de CREC revisa cada paquete de extracción para comprobar que los términos y los equivalentes validados son correctos.

Análisis de datos y evaluación de resultados.

En el último paso, que hemos realizado los colaboradores de CREC, se han tenido en cuenta los siguientes datos:

- cantidad de términos extraídos
- cantidad de términos validados
- cantidad de equivalencias extraídas
- cantidad de equivalencias validadas
- problemas de ruido o silencio producidos en la extracción

RESULTADOS

En este apartado mostramos dos tablas con los resultados de los diez primeros proyectos de extracción automática llevados a cabo en el proyecto CREC durante el curso académico 2008-2009.

La primera tabla se corresponde con los términos extraídos en español. En la primera columna aparece el nombre de cada paquete. Los primeros cinco paquetes tienen un menor número de palabras que los últimos cinco, aproximadamente tres veces menos. Esto se debe a que los primeros paquetes de alineación se dividieron en tres partes para limitar el número de términos extraídos automáticamente.

En la segunda columna se observa el número de términos validados manualmente por los estudiantes con respecto a los candidatos propuestos por el programa. Por último, se puede ver que el porcentaje de términos validados en español es de 21,17 %.

Paquete	Términos validados/candidatos a término
VBce01-2	105/465
VBce01-3	70/275
VBce02-1	132/366
VBce02-2	193/354
VBce02-3	69/300
VBce03	220/1220
VBce04	184/920
VBce05	143/1026
VBce06	166/1206
VBce07	256/1127
% Términos validados	21’17 %

La segunda tabla muestra los resultados relativos a las equivalencias validadas por los estudiantes respecto a las propuestas por el programa. El porcentaje de equivalencias validadas es de 9,4%.

Paquete	Equivalencias validadas/candidatos a
VBce01-2	109/1024
VBce01-3	70/562
VBce02-1	132/760
VBce02-2	197/761
VBce02-3	69/680
VBce03	221/2830
VBce04	187/2182
VBce05	152/2483
VBce06	169/2847
VBce07	266/2568
% Equivalencias validadas	9,4%

DISCUSIÓN

La conclusión principal que se deriva de estos resultados es la gran cantidad de ruido que el programa ha generado en su extracción automática tanto en la extracción de términos como en las equivalencias propuestas. Solo el 21,17% de los candidatos extraídos han sido aceptados como términos en la validación manual. En el caso de las equivalencias, el porcentaje es aún menor, sólo el 9,4% de las equivalencias propuestas por el programa se han considerado equivalentes válidos.

Como se puede ver en la tabla de resultados, en los paquetes con menor número de palabras (VBce01-1, VBce01-2, VBce01-3, VBce02-1, VBce02-2, VBce02-3) el ruido es menor que en los paquetes más grandes (VBce03, VBce04, VBce05, VBce06, VBce07).

A continuación presentamos brevemente los problemas detectados en la extracción de terminología que pueden ser considerados como el origen del ruido producido. Presentamos en primer lugar los errores detectados en la extracción de terminología y después los problemas en la asignación de equivalencias.

Extracción de nombres propios
En el corpus aparecen gran cantidad de nombres propios referentes a empresas de producción cerámica o ciudades en las que se produce. Estos nombres no son relevantes para un recurso terminológico. Sin embargo, el programa de extracción los propone como candidatos.

Extracción de unidades no lingüísticas
El programa propone entre los candidatos a término unidades no lingüísticas como, por ejemplo, números o símbolos.

Extracción de unidades no especializadas
ExtraTerm ha propuesto una larga lista de palabras pertenecientes al lenguaje general como, por ejemplo, *utilizar, producir, trabajar, etc.* Estas unidades no especializadas tienen una alta frecuencia en el corpus, pero no son términos válidos del campo de la cerámica.

Extracción de variantes morfológicas
Entre las listas de candidatos a término podemos encontrar gran número de formas verbales conjugadas (por ejemplo, *crea, fabrica, etc.)* y sustantivos en plural (*baldosas, piezas).* Estas unidades deberían aparecer una única vez en su forma canónica, es decir, sin morfemas flexivos.

Extracción de candidatos a término demasiado largos

En numerosos casos, los candidatos a término consisten en términos válidos rodeados de unidades no terminológicas como, por ejemplo, *curvas de reflactancia obtenidas a partir, circón y dio lugar a vidriados.*

Extracción fragmentada del término

En algunos casos el programa ha extraído únicamente uno de los elementos que conforman un término, por ejemplo, *supensión* en lugar de *suspensión acuosa.* Una de las causas puede ser el llamado silencio intrínseco que tiene lugar cuando por razones discursivas se elimina una parte del término (a este fenómeno también se le llama deixis), debido a que el término completo aparece anterior o posteriormente.

A continuación presentamos los problemas detectados en relación a las equivalencias propuestas por el programa.

Equivalencia errónea

En numerosas ocasiones el programa ofrece candidatos que no son equivalentes como, por ejemplo, *deslizamiento* y *resistance, filtro* y *solid,* etc.

Extracción fragmentada de los equivalentes.

Al igual que con los términos en español, el programa también extrae algunas equivalencias fragmentadas. Por ejemplo, para el término *resistencia al deslizamiento del calzado,* el programa propone como equivalente *slip resistance,* en lugar de *footware slip resistance.*

Candidato a equivalente demasiado largo.

Frecuentemente el programa extrae candidatos a equivalente demasiado largos, que se corresponden con la suma de dos términos o con la suma de un término y de unidades no terminológicas. Por ejemplo, para el término *temperatura de reblandecimiento,* el programa propone como equivalente *softening temperature and glass transformation temperature.* Asimismo, para el término *centro de innovación y tecnología* el programa propone como equivalente *co-operation with a centre for innovation and technology.*

Ausencia de equivalente

En ocasiones el programa no propone ningún equivalente para el término en español, aunque éste sí que se encuentre traducido en el corpus. Por ejemplo, el programa no ofrece equivalente para el término *banda de devitrificación.*

CONCLUSIÓN

Con los resultados de este estudio podemos concluir que la herramienta de extracción automática Multiterm 2007 Extract ayuda a agilizar las tareas de extracción de terminología. Esta extracción automática facilita la elaboración de bases de datos bilingües que se pueden integrar en las herramientas de traducción asistida o que sirven como recurso terminológico para la traducción especializada. Sin embargo, los problemas que hemos detectado en nuestro análisis muestran que las herramientas de extracción todavía presentan carencias que implican mucho trabajo manual por parte del terminólogo. En la validación manual es necesario recurrir a los contextos del corpus para comprobar o identificar términos y equivalentes que el programa no ha extraído correctamente.

BIBLIOGRAFÍA

- Alcina, A. (2003): «La programación de objetivos didácticos en Terminótica atendiendo a las nuevas herramientas y recursos», en Gallardo, N. (ed.): *Terminología y traducción: un bosquejo de su evolución*, Granada, Atrio, 79-91.
- Alcina, A., V. Soler, N. Vidal y C. Martí (2007): «CREC II: Creación de recursos lingüísticos electrónicos para la Informática aplicada a la traducción y Terminología. Una experiencia de trabajo en equipo con estudiantes», en *VII Jornada de Millora Educativa*, Unitat de Suport Educatiu, Universitat Jaume I, Castellón.
- Benavent, P. y S. Parrilla (2007): «Análisis de la extracción automática de términos con el programa informático ExtraTerm», *Fòrum de Recerca: XII Jornadas de Fomento de la investigación en Traducción*, Número 12, Universitat Jaume I, Castellón.
- Déleger, L. M. Merkel y P. Zweigenbaum (2006): «Using word alignment to extend multilingual medical terminologies», en *Proceedings of LREC 2006 Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine*, Génova, Italia.
- Estopà, R. (1999): «Eficiencia en la extracción automática de terminología», *Perspectives: Studies in Traductology*, 7(2), 277-286.
- Kraif, O. (2008): «Extraction automatique de lexique bilingue: application pour la recherche d'exemples en lexicographie», en Maniez, F., P. Dury, N. Arlin y C. Rougemont (eds.): *Corpus et dictionnaires de langues de spécialité*, Presses Universitaires de Grenoble, Grenoble, 67-87.
- L'Homme, M.C. (2005): «Sur la notion de 'terme'», *Meta* 50(4), 1112-1132.
- McEnery, A., J.-M. Langé, M. Oakes y J. Véronis (1997): «The exploitation of multilingual annotated corpora for term extraction», en Garside, R., G. Leech y A. McEnery (eds.): *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Addison Wesley Longman, London, 220-230.

- Soler, V., A. Alcina y A. Estellés (2006): «La digitalización de textos para la elaboración de un corpus lingüístico electrónico: una experiencia de trabajo en equipo con estudiantes», en *X Jornades de Traducció i Interpretació a Vic: Tecnologies a l□abast*, Universitat de Vic.
- Vivaldi Palatresi, J. y H. Rodríguez Hontoria (2007): «Evaluation of terms and term extraction systems. A practical approach», *Terminology* 13(2), 225-248.