독일언어문학 제68집(2015. 6). 21-42

‘AntConc’를 활용한 독일어 전문용어후보 자동추출방안

홍문표 (성균관대)

I. 들어가는 말

한 언어에 아무리 능통한 번역가나 통역가라 할지라도 일반분야가 아닌 전문분야의 통번역은 상당히 난이도가 높은 작업이다. 전문분야의 통번역이 어려운 가장 큰 이유는 해당 분야에 등장하는 전문용어의 통번역이 어렵기 때문이다. 해당 분야에 대한 지식이 부족한 경우 언어지식만을 사용한 번역은 어색한 번역, 심지어는 오역까지도 초래할 수 있다(Vgl. 홍문표 2008).

번역작업을 위해 비교적 긴 시간이 주어지는 번역프로젝트의 수행 시에는 해당 전문분야에 출현하는 전문용어의 데이터베이스를 미리 구축해놓고 번역을 수행할 수 있다. 물론 이 경우에도 전문용어 데이터베이스를 어떻게 구축하는가는 또 다른 큰 문제이지만 번역가의 입장에서는 상대적으로 수월하게 번역작업을 수행할 수 있다. 그러나 통번역을 위한 준비시간이 매우 짧게 주어지는 시나리오에서는 전문용어 데이터베이스를 짧은 시간 내에 어떻게 구축할 것인가가 매우 중요한 문제이다.

예를 들어 전문분야에 대한 독일어 통역의뢰를 받은 통역가의 경우 전문용어의 통역에 대한 심리적 부담감이 매우 클 것이다. 이 경우 해당분야 코퍼스 내지는 텍스트로부터 짧은 시간 내에 간편하게 전문용어를 추출하여 해당 분야 전문가에게 용어번역결과의 감수를 받는다면 훨씬 수월하게 통역업무를 수행할 수 있을 것이다.

본 연구는 최근 통계기반의 언어학연구를 위한 보조도구로써 널리 활용되고 있는 ‘AntConc’를 활용하여 독일어 텍스트로부터 전문용어 후보를 추출하는 방법론을 다룬다. ‘AntConc’는 전문용어 추출을 위한 ‘SDL MultiTerm’ 등과는 달리 프리웨어 Freeware이므로 누구나 무료로 쉽게 사용할 수 있다. 따라서 상대적으로 영세한 통번역 환경에 있는 한국어-독일어 통번역가들이 쉽게 접근할 수 있는

장점이 있다.

본 논문의 구성은 다음과 같다. 2장에서는 통번역 작업을 위한 전문용어의 중요성에 대해 다룬다. 여기서는 전문용어의 정의와 전문용어후보에 대한 정의가 내려질 것이다. 또한 실제 통번역 작업에서 전문용어의 처리가 어느 정도의 문제를 일으키는지도 살펴볼 것이다. 3장에서는 전문용어후보의 자동추출을 위한 기존연구를 살펴볼 것이다. 전문용어후보의 자동추출 방법은 언어학적 방법론과 통계적 방법론 등으로 나뉠 수 있는데, 각 방법론의 장단점을 보게 될 것이다. 4장에서는 본 연구에서 제안하는 방법론을 자세히 설명한다. 우선 'AntConc' 툴에 대한 간단한 소개와 함께 전문용어 추출을 위한 기능을 살펴본다. 'AntConc'에는 전문용어 추출을 위한 기능이 따로 마련되어 있지는 않으나 키워드 추출 기능을 적절히 활용할 경우 상당히 유용한 전문용어후보 추출기능으로 작동할 수 있음을 보일 것이다. 5장에서는 본 연구에서 제안한 방법론의 성능을 살펴보기 위한 실험을 수행한다. 이 실험에서는 키워드 Schlüsselwort 추출에 사용되는 레퍼런스 코퍼스 Referenzkorpus와 레퍼런스 단어리스트 Referenzwörterlist의 역할에 대한 논의가 이루어질 것이다. 마지막으로 6장에서는 연구결과를 정리하고 향후 연구방향을 제시하게 될 것이다.

II. 번역과 전문용어

II.1. 전문용어의 개념

본 연구에서 언급하는 전문용어 Term은 기술분야 등과 같은 특정 주제영역에서 특수한 개념을 지시하는 것으로 분야전문가에 의해 판정되는 명사나 명사구를 가르킨다(Vgl. Jacquemin/Bourigault 2000). 김성원/김정우(2011)의 정의에 따르면 전문용어란 일상생활에서 사용되는 일반어가 아니라 사용자가 해당분야의 전문인들로 한정되어 있는 용어를 말한다.

이들의 연구에 따르면 전문용어는 다의성을 갖는 일반용어와는 달리 개념과 일대일의 관계를 갖는다. 전문용어에 대한 이들의 입장은 용어의 사용분야 뿐만

아니라 용어의 사용자가 일반인이냐 전문인에 한정되는가에 달려있다. 본 연구에서는 이러한 입장을 수용하여 어떠한 용어가 분야 전문가 뿐만 아니라 일반인들도 널리 사용하는 경우에는 전문용어로 보지 않고 일반용어로 분류한다.

이러한 정의에 따르면 예를 들어 자동차기술 분야의 독일어 전문용어를 선정할 때 아래 (1)과 같은 용어들은 전문용어로 분류가 가능하겠지만, (2)의 단어들은 자동차기술 분야에서도 사용되지만 대부분의 일반인들도 일상생활에서 자주 사용하는 용어이므로 전문용어로 분류할 필요는 없어 보인다.

(1) *Einkolbenpumpe, Einlassnockenwellenverstellung, Einspritzmenge, Einstellbuchse, Geblaseabsenkung, Gepäckraumabdeckung, Geschwindigkeitsregelanlage, Giermomente, Handschaltniveau, Heckklappenverkleidung, Heckscheibenrollo, Hochspannungkabel, Klemmhebel, Kraftstoffabschaltung, Kupplungsausrückbetätigung, Lambdasonde, Luftwiderstandsbeiwert, ...*

(2) *Airbag, Auto, Batterie, Benzin, Diesel, Fahrzeug, Kraftstoff, Motor, Reifen, Reparatur, ...*

전문용어와 일반단어를 구분짓는 또 하나의 큰 기준은 분야전문가에 의한 채택여부이다. Jacquemin/Bourigault(2000)의 주장처럼 전문분야 문서에 등장하는 단어들은 우선적으로는 전문용어후보 Termkandidat이고 분야전문가들에 의해 채택되면 비로소 전문용어로 인정된다. 따라서 본 연구에서는 전문분야 텍스트로부터 자동으로 추출하고자 하는 대상을 전문용어후보라고 간주한다.

II.2. 전문용어의 번역문제

김정우(2003)는 자연과학텍스트의 번역문제를 연구하였는데, 특히 텍스트의 특성상 그 목적이 지식의 전달에 있으므로 번역자의 개성을 최대한 억제하고 객관적인 관점에서 정확하게 번역을 해야 한다고 보았다. 이를 위해 특히 전문용어의 번역이 중요한데 다음과 같은 영어문장의 예를 들어 분야별 전문용어 번역의 중요성을 지적한다.

(3) a. It makes a noise like a train
열차같은 소음을 낸다
b. Noise is always present in an amplifier
증폭기에는 반드시 잡음이 있다

(4) a. The function of the heart is to pump the blood through the body
심장의 역할은 몸에 혈액을 펌프작용으로 보내는 것이다
b. For example, cos x is an even function
예를 들면 코사인 x는 우함수이다

위의 예문을 보면 (3)a.의 경우 'noise'의 번역이 기계공학이나 환경공학 등의 영역에서는 '소음'으로 되고, (3)b.의 경우에는 전기공학이나 전자공학 등에서는 '잡음'이 더 적합하다. 유사하게 (4)a.의 경우 'function'의 번역이 의학 분야 등에서는 '역할'로 될 가능성이 높지만, (4)b.의 경우에서처럼 전자공학이나 수학 등의 분야에서는 '함수'로 될 가능성이 높으며, 만약에 이 경우 '역할'이라고 번역하면 이는 잘못된 번역이라 할 수 있다.

방교영(2000)은 국제회의 통역에서 전문용어의 순발력 있는 구사를 위해 자료수집 및 전문용어의 준비를 강조하고 있다. 통번역의 준비과정에서 전문분야에 대한 지식을 쌓는 것 만큼 분야별 전문용어의 준비는 통번역의 질을 결정하는 중요한 요소가 된다. 그에 따르면 성공적인 통번역 업무의 수행을 위해서는 제한된 시간 내에 효율적으로 전문용어를 습득하는 것이 중요하다.

II.3. 독일어 전문용어의 특징

Hong et al. (2001)의 연구에서는 기술분야 독일어 전문용어의 언어학적 특징을 다음과 같이 제시하였다. 첫째, 독일어 전문용어는 영어와 유사하게 합성어가 많지만 영어와는 달리 단어 중간에 공백을 허용하지 않는다. 위 연구에서 발췌한 다음의 예를 보자.

(5) Schwingungskomfort : Vibration comfort

Wankstabilisierung : anti-roll stabilization
Alu-Hohlspeichenfelgen : Hollow spoked aluminum wheel
Gas-Generator : Gas generator

(5)의 예에서 보는 바와 같이 영어에서는 합성어를 구성하는 성분들이 공백에 의해 나뉘어 있지만 독일어에서는 합성명사가 연결사 '-s'또는 연결부호 '-' 등을 통해 표기되므로 공백이 등장하지 않는다 (Vgl. Hansen-Schirra et al. 2007). Hong et al. (2001)의 연구에서는 독일어 기술문서에서 주제영역에 따라 복합명사 전문용어의 비율이 약 57%에서 최대 94%까지 차지한다고 밝혔다. 이러한 형태론적인 특성은 독일어 전문용어 후보의 추출에 고려할 수 있다.

두 번째 특징은 영어단어로부터 그대로 차용된 용어가 많다는 점이다. 예를 들어 자동차 기술 분야에서 'Distronic', 'Power-Program', 'Economy-Program' 등은 영어에서 그대로 차용된 전문용어들이다.

마지막으로는 명사구 형태로 된 전문용어가 전체 전문용어 중 분야에 따라 최대 45% 정도까지 육박하는 경우가 있다. 이 경우 통사패턴은 '형용사+명사', '명사구+명사구', '명사+전치사구' 등의 구조이며 전치사구는 주로 전치사 'mit'를 핵심어로 하는 경우가 많다. Hong et al. (2001)에서 가져온 다음 (6)의 예는 명사구 형태로 되어 있는 자동차 기술분야의 명사구 전문용어들이다.

(6) ADJ + N : induktive Antenne, passiver Gasdruck-Stoßdämpfer, dreischalige A-Säule
N + NP_{GEN} : Armauflage der Mittelkonsole, Wippbewegung der Wippspitze
N + PP : Schiebedach mit Memoryfunktion, Automatikgetriebe mit fünf Fahrstufen

III. 전문용어후보 추출에 관한 기존연구

III.1. 형태-통사지식에 기반한 추출방법

텍스트상에 등장하는 단어가 전문용어후보인지의 여부를 판단하는 가장 간단한 기준은 후보단어의 형태론적 특성을 파악하는 것이다. Heid(1998, 1999)는 입

력텍스트의 품사 태깅 단계 이후에 단어의 접사정보를 활용하여 전문용어후보를 추출하고자 하였다. 그는 입력문장에 대해 정규식표현 Regulärer Ausdruck을 적용하여, 다음과 같은 접두사 (7)과 접미사 (8)이 부착된 단어가 기술문서상에 출현하면 이 단어는 전문용어후보일 가능성이 높은 것으로 파악하였다.

(7) ab.+, auf.+, ent.+, anti.+, bi.+, mega.+, mikro.+, multi.+, radial.+, semi.+, ad.+, ex.+, in.+, ko.+, pro.+

(8) .+grad, .+heit, .+nis, .+schaft, .+tum, .+ial, .+gramm, .+graph, .+id, .+ik, .+tion, .+tät, .+um, .+ator.

(7), (8)번의 접사 Affix들 중 일부는 엄밀히 말하면 형태론적인 접사라고 보기 힘든 것들이 있으나, 전문용어후보의 추출을 위해 위와 같은 스트링 Zeichenkette 정보를 활용하면 도움이 될 수 있다고 주장하였다.

그 밖에 구구조 Phrasenstruktur로 되어 있는 전문용어후보를 탐색하기 위해 품사태그정보도 활용할 수 있다. Heid(1998)는 특히 N+N, N+PP, A+N, N+A, N+ADV 품사열에 주목하였는데, 이와 같은 품사열로 구성된 표현을 문서내에서 모두 추출한 후 도메인에 특화된 형태소가 등장할 경우 해당 표현을 전문용어후보로 간주하였다. 도메인 특화 형태소는 코퍼스 분석을 통해 상대빈도를 이용해 구축한다. 자동차 기술분야의 예에서 그가 도메인 특화 형태소로 제시한 형태소들의 예는 다음과 같다.

(9) -fahr-, -motor-, -trieb-, -bau-, -stoff-, -elektr-, -system-, -auto-, -techn-, -filt-, -kanal-, -tank-, ..., -park-

Hong et al. (2001)의 연구에서는 입력문장을 품사태깅 단계를 거친 후 특정 품사열에 대해 정규식 표현을 적용하여 추출하였다. 이들의 연구에서는 특히 'ADJ+N'으로 구성된 후보들에 제외단어 리스트 stop words list를 적용하여 후보들을 필터링하였다. 일반적으로 제거단어 리스트는 관사와 조동사 등과 같은 기능어들이지만 이들의 연구에서는 전문용어에 일반적으로 사용되지 않는 형용사 리스트 (10)을 작성하여 제외단어로 취급하였다는 점에서 기존의 연구와 차별되

는 점이 있다.

(10) ander, außergewöhnlich, beachtlich, beide, besonder, bisherig, deutlich, echt, eigen, einfach, einzig, erforderlich, erst, ganz, gemeinsam, genau, gering, gesamt, gleich, herkömmlich, hochwertig, jeweilig, komplett, konkret, konventionell, kostenlos, lebenslang, luxuriös, maßgeblich, möglich, neu, notwendig, perfekt, richtig, schließlich, sinnvoll, sogenannt, solch, speziell, spezifisch, tatsächlich, teilweise, traditionsreich, typisch, üblich, übrig, unterschiedlich, ursprünglich, viel, vorbildlich, weiter, wenig, wichtig, zahlreich, …

III.2. 통계지식에 기반한 추출방법

전문용어후보를 자동으로 추출하는 가장 대표적인 방법은 통계지식을 활용하는 것이다. 이 방법의 기본 아이디어는 어떠한 전문분야문서의 내용을 가장 잘 특징짓는 단어는 그 문서에는 자주 출현하지만 다른 일반분야문서에서는 자주 출현하지 않을 것이다라는 생각이다(Vgl. Witschel 2005). 즉, 'der', 'ein', 'für' 등과 같은 기능어들은 기술문서에도 자주 출현하겠지만 일반 문서에도 자주 등장하게 되므로 이러한 단어는 전문용어후보에서 배제될 것이다. 이러한 아이디어를 가장 쉽게 구현한 방법론이 'TF/IDF (term frequency/inverse document frequency)'이다.

문서의 토픽을 구분하는데 일반적으로 사용되는 이 개념을 비슷한 방식으로 전문용어후보의 추출에도 사용할 수 있다. Ahmad et al. (1992)의 연구에서 제안된 것과 같이 어떤 단어가 전문용어후보로 간주될 수 있는지의 여부는 해당 단어가 전문분야 텍스트에 출현하는 빈도를 일반분야 코퍼스에서 출현하는 빈도와 비교한 상대빈도가 중요하다.

로그우도비 Log-likelihood는 단어들 간의 긴밀도를 계산하는데도 성공적으로 활용될 수 있는데, Hong et al. (2001)의 연구에서는 구문분석 결과로 추출된 바이그램의 로그우도비를 계산하여 약 75%의 추출정확률을 달성할 수 있음을 보였다.

Ⅳ. 제안하는 방법론

Ⅳ.1. AntConc

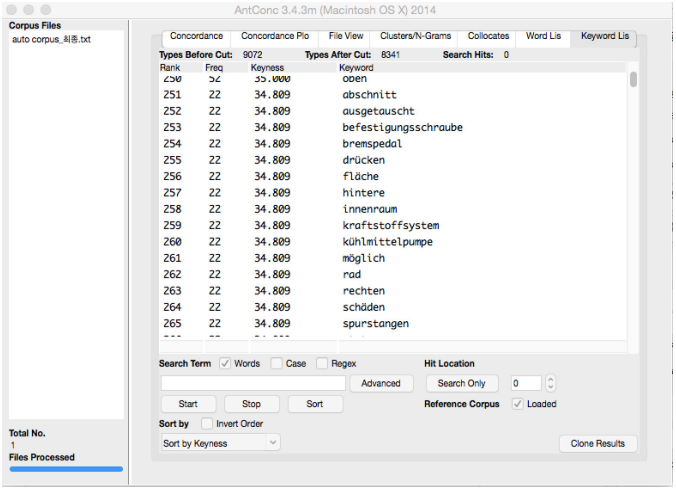
본 연구에서는 누구나 인터넷에서 무료로 손쉽게 다운받을 수 있는 'AntConc' 프로그램이 상용 전문용어추출 프로그램과 비교하여 전혀 손색없이 사용될 수 있음을 보이고자 한다. 'AntConc'를 전문용어후보 추출을 위해 사용하는 방법을 소개하기 전에 우선 이 프로그램에 대한 전반적인 소개를 하고자 한다.

'AntConc'는 일본 와세다 대학의 Laurence Anthony 교수에 의해 개발되어 현재 윈도우용으로는 버전 3.4.3까지 소개되어 있으며 맥 OS 및 리눅스도 지원한다(Vgl. L. Anthony 2014). 'AntConc'의 주요 기능으로는 코퍼스에 출현하는 단어 리스팅 기능, 콘코던스 Konkordanz 분석, n그램 분석, 연어 Kollokation분석, 키워드 분석 등이 있다.

단어 리스팅 기능은 코퍼스에 등장하는 전체 토큰의 개수, 타입의 개수를 알려주며 모든 출현하는 토큰을 빈도수와 함께 제시해준다. 단어 리스팅은 코퍼스에 출현한 모든 단어, 좀 더 정확히 말하자면 토큰을 제시해주므로 우리가 찾고자 하는 전문용어후보들도 이 리스트 안에 들어있다고 볼 수 있다. 그러나 이 단어리스트에는 전문용어 뿐만 아니라 코퍼스에 등장하는 모든 단어가 수록되어 있으므로 이 단어리스트로부터 좀 더 전문용어에 가까운 전문용어후보들을 추출할 필요가 있다.

우리가 여기서 전문용어후보 추출을 위해 좀 더 자세히 살펴보고자 하는 기능은 키워드분석 기능이다. 'AntConc'는 어떤 문서의 키워드를 찾을 수 있는 기능을 제공하는데 여기에는 크게 두 가지의 방법이 있다. 첫 번째는 레퍼런스 코퍼스를 이용하는 방법이고 두 번째는 단어리스트를 이용하는 방법이다.

그림 1은 레퍼런스 코퍼스를 이용하여 자동차기술 분야의 문서로부터 키워드를 추출해내는 스크린샷이다.



<그림 1> ‘AntConc’의 키워드추출 기능

다음 절에서는 키워드 추출 기능을 전문용어 후보추출을 위해 활용하는 방법론에 대해 소개한다.

Ⅳ.2. 키워드 추출

전문용어 후보를 추출하는데 키워드 개념을 사용하는 이유는 전문용어 후보는 일반텍스트에서 보다는 전문분야 텍스트에서 상대적으로 빈번하게 등장할 것이라는 예측에 기인한다. Knorz(1991)는 색인용어 index term을 다음과 같이 정의하였다.

“Wort, das den Inhalt eines Dokumentes kennzeichnet.”

이 정의에 따르면 색인용어는 어떤 문서의 내용을 가장 잘 드러내는 것이고 이것은 동시에 해당 분야의 전문용어이다. 따라서 키워드 혹은 인덱스 용어를 추출하는 방법이 전문용어 후보를 추출하는 방법과 다르지 않다.

이러한 점을 고려할 때 상대빈도수 Relativfrequenz는 전문용어후보를 결정하는 좋은 기준이 될 것이다. Rayson/Garside(2000)는 어떤 도메인의 텍스트를 다른 도메인의 텍스트와 구별시켜주는 대표적인 단어를 찾기 위해 로그우도비검증 log-likelihood ratio test를 적용하였다. 로그우도비검증 수식은 어떤 단어가 두 개의 코퍼스에 등장할 때 상대빈도가 큰 경우 큰 값을 갖게 된다. 또한 어떤 단어가 도메인에 상관없이 고른 출현빈도수를 보인다면 이 단어의 로그우도비검증 값은 낮은 값을 갖게 된다. 로그우도비는 아래 표1과 같은 분할표 contingency table을 작성하여 얻게 된다.

<표 1> 로그우도비 계산을 위한 분할표

	코퍼스1	코퍼스2	합계
단어 빈도수	a	b	a+b
기타 모든 단어빈도수	c-a	d-b	c+d-a-b
합계	c	d	c+d

위에서 c값은 코퍼스 1에 등장하는 모든 단어의 총 빈도수이고 d값은 코퍼스 2에 등장하는 모든 단어의 총 빈도수이다. a값은 어떤 특정한 단어가 코퍼스 1에서 출현한 빈도이고, b값은 그 단어가 코퍼스 2에서 출현한 빈도이다. 로그우도비는 위와 같은 분할표상의 값으로부터 기댓값 expected values를 구한 후 그 기댓값을 사용하여 구해진다.

기댓값은 다음과 같은 수식으로 구해지는데 이 수식에서는 두 개의 코퍼스의 크기가 고려되므로 별다른 정규화의 필요가 없다는 장점이 있다. 이 수식에서 N은 두 개의 코퍼스에 등장하는 모든 단어의 빈도수이고, O는 분할표상의 a와 b, 즉 특정단어가 각각의 코퍼스상에 등장한 빈도수이다.

(11)
$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

여기서 N1은 c이고 N2는 d가 되므로 위의 경우에서 E1=c*(a+b)/(c+d), E2=d*(a+b)/(c+d)이다. 이를 이용하여 로그우도비를 구하는 공식은 다음과 같다.

(12) LL = 2 * ((a * log(a/E1)) + (b * log(b/E2)))

위 수식은 두 기댓값의 차이가 클수록 큰 값을 갖게 된다. 즉, 다시 말하면 로그우도비의 값이 클수록 어떤 단어가 두 집단에 출현하는 빈도가 상대적으로 큰 차이가 난다는 의미이고, 이것을 우리의 목적에 적용하면 어떤 분야 문서의 키워드, 즉 전문용어후보일 가능성이 높다는 것을 의미한다.

계산된 로그우도비의 p-값 p-value이 통계적으로 유의미한 범위인 0.05 미만이 되기 위해서는 로그우도비의 값이 3.64이상이 되어야 한다. 즉, 본 연구에서는 로그우도비의 값이 3.64이상인 단어를 전문용어후보로 받아들일 때 그 정확도가 어느정도인지 보게 될 것이다.

'AntConc'는 어떤 분야의 텍스트에서 키워드를 추출하기 위한 방안으로 레퍼런스 코퍼스를 사용하는 방법과 키워드 리스트를 사용하여 로그우도비를 계산하는 두 가지의 방법을 제공하는데, 본 연구에서는 두 가지의 방법을 모두 시도하였고 그 성능의 차이를 비교하였다.

Ⅴ. 실험

Ⅴ.1. 실험준비

본 연구에서 제안하는 방법론을 실험하기 위해 자동차분야의 기술문서를 수집하였다. 문서내용의 전문성을 갖추기 위해 단순히 사용자를 위한 사용설명서가 아닌 기술자들을 위한 정비매뉴얼을 중심으로 코퍼스를 구축하였다.[1] 자동차

1) 자동차 정비매뉴얼의 출처는 다음과 같다:
- Reparaturleitfaden Audi A3 2004
- Reparaturleitfaden Audi A4 2001

분야 코퍼스는 총 68,938개의 단어토큰으로 이루어져 있으며 총 9,072개의 단어유형이 사용되었다.

이와 비교하기 위한 레퍼런스 코퍼스는 일반분야 텍스트로 구성되어 있으며 경제, 사회, 문화, 스포츠 등의 주제로 이루어져 있다. 이 코퍼스는 83,128개의 단어토큰으로 이루어졌으며 총 10,394개의 단어유형이 사용되었다. 일반적으로 코퍼스가 레퍼런스 역할을 하기 위해서는 충분히 많은 어휘와 언어현상을 포함할 수 있도록 다양한 분야, 다양한 장르 등의 출처로부터 구축되어야 한다. 이를 통해 코퍼스의 대표성이 보장될 수 있다.

이러한 레퍼런스의 목적으로 사실상 가장 적합한 코퍼스는 독일 IDS의 'DeReKo' 코퍼스라고 할 수 있는데, 이 코퍼스는 그 규모가 방대하여 'COSMAS II'라고 하는 인터페이스만을 통해 데이터베이스에 접근할 수 있다(Vgl. Kupietz et al. 2010). 'AntConc'는 텍스트 파일 형태의 레퍼런스 코퍼스를 로딩하여 사용하므로 'DeReKo'를 사용하는 것은 현실적으로 불가능하다.

따라서 본 연구에서는 자체적으로 구축한 레퍼런스 코퍼스를 사용하였는데 레퍼런스 코퍼스의 충실성에 따라 키워드 추출의 결과가 달라지므로 향후에는 좀 더 대표성이 확보된 코퍼스를 사용할 필요가 있다.

키워드를 추출하는 두 번째 방법은 일반단어리스트와 비교하는 것이다. 일반 독일어 코퍼스로부터 빈출하는 순서대로 추출된 단어리스트가 있다면 이 단어들은 전문분야 텍스트에만 빈번하게 등장하는 전문용어후보를 찾아내기 위한 좋은 비교자료로 활용될 수 있을 것이다.

본 연구에서는 라이프치히 대학에서 구축한 독일어 빈출단어리스트를 활용하였다.[2] 이 단어들은 일반분야 독일어 텍스트에 가장 빈번히 등장하는 독일어 단어들로서 렘마 lemma 형태가 아닌 단어형태 Wortform를 기준으로 배열되어 있다. 즉, 'liebt'와 'lieben'은 서로 다른 엔트리로서 등재되어 있다. 해당 웹사이트

- Ford FAHRZEUGE VAZ-2110, VAZ-2111, VAZ-2112 REPARATURANLEITUNG
- AUDI A2 – Technik Konstruktion und Funktion Selbststudienprogramm 240
- AUDI A2 - Motor und Getriebe Konstruktion und Funktion Selbststudienprogramm 247
- VW Selbststudienprogramm 339 Der Passat 2006

2) URL: wortschatz.uni-leipzig.de/html/wliste.html

에서는 빈출단어 리스트를 100단어, 1,000단어, 10,000단어 단위로 제공하는데 본 연구에서는 10,000단어 리스트를 사용하였다. 참고로 위 리스트에서 가장 순위가 높은 상위 30개의 독일어 단어는 다음과 같다.

<표 2> 라이프치히 대학 단어리스트 상위 30단어

	빈도순 단어
1~10위	der, die, und, in, den, von, zu, das, mit, sich
11~20위	des, auf, für, ist, im, dem, nicht, ein, Die, eine
21~30위	als, auch, es, an, werden, aus, er, hat, dass, sie

라이프치히 대학의 단어리스트 이외에도 사용가능한 레퍼런스 단어리스트로는 IDS에서 제공하는 'DeReWo'가 있으나 라이프치히 대학의 단어리스트와 순위가 크게 다르지 않다.[3] 그러나 'DeReWo'는 2012년 12월에 발표된 최신버전의 경우 326,949개의 단어기본형을 수록하고 있다. 이토록 큰 엔트리의 수가 전문분야 문서로부터 전문용어후보를 추출하는데 도움이 될지 아닐지는 아직 검증되지 않았으므로 본 연구에서는 1만 단어 규모의 라이프치히 대학의 단어리스트를 사용하였다.

그러나 라이프치히 대학의 단어리스트는 빈도순으로 단어의 순위가 매겨져있을 뿐 해당 단어의 빈도가 수치로 나타나 있지는 않다. 'AntConc'의 키워드 추출 기능을 사용하기 위해서는 단어리스트에 빈도값이 함께 기재되어 있어야 하므로 본 연구에서는 임의로 가장 순위가 높은 단어의 빈도값을 10,000으로 하고 2위는 9,999, 3위는 9,998과 같이 1씩 줄여서 최종적으로 10,000번째 단어는 빈도수가 1이 되도록 하였다.[4] 이러한 임의의 빈도수 조정이 키워드 추출에 있어서 어떠한 심각한 영향을 미치는지도 중요한 연구주제이긴 하나, 본 연구에서는 다

3) URL: http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html

4) 익명의 심사자께서 지적하신 바와 같이 단어의 빈도수를 임의로 조정하는 것은 실험결과에 큰 영향을 끼칠 수 있으므로 신중히 다루어야 할 부분이다. 그러나 본 연구의 가장 큰 목적은 'AntConc'를 사용한 간단한 방법만으로도 상용시스템에 못지 않은 전문용어후보 추출 성능을 올릴 수 있다는 것을 보이는 것이기 때문에 본 연구에서는 이에 대해 본격적으로 다루지는 않는다.

루지 않았고 향후 연구테마로 다루기로 한다.

V.2. 실험결과

본 연구에서 제안하는 방법론을 검증하기 위해 기술문서로부터 자동으로 전문용어후보를 추출하는 정확률을 측정하였다. 비교대상은 'AntConc'의 키워드 추출기능 중 레퍼런스 코퍼스를 사용하여 추출하는 방법과 라이프치히 대학의 단어 리스트를 사용하여 추출하는 방법이었다. 실험 대상은 앞서 소개한 자동차 기술분야의 기술문서였다. 본 실험에서는 두 방법 모두에서 로그우도비가 3.64 즉, p-값이 0.05 미만인 단어들만을 고려하였다.

각 방법론에 의해 자동으로 추출된 키워드들은 두 명의 독일어 전문가에 의해 검증되었다. 두 명의 독일어 전문가가 모두 전문용어후보로 인정하는 단어만 전문용어후보로 간주하였고 나머지 단어들은 비전문용어로 간주하였다.[5] 본 연구에서는 위와 같이 추출된 키워드 중 각 방법에서 상위 2천개의 키워드만을 대상으로 정확률을 검토해보았다. 두 방법 모두 상위 2천개의 키워드는 로그우도비가 최하 4 이상으로 p-값이 0.05 미만의 신뢰도가 높은 단어들이었다.

추출된 키워드는 미리 컴파일해 놓은 제외단어 리스트와 비교해보고 이 리스트에 들어있는 단어가 키워드에 속해 있으면 이들 단어는 제거하였다. 제외단어 리스트는 단일 단어로서 절대 전문용어가 될 수 없는 단어들인 기능어들로 구성되어 있으며, 여기에는 주로 관사, 전치사, 조동사 등이 속한다.

레퍼런스 코퍼스를 사용하여 추출한 2천개의 키워드에는 총 35개의 제외단어가 포함되어 있었으므로 실질적으로는 1,965개의 키워드가 추출되었다. 단어 리스트를 사용하여 추출한 2천개의 키워드에는 총 55개의 제외단어가 포함되어 있었기 때문에 최종적으로는 1,945개의 키워드가 추출되었다.

레퍼런스 코퍼스 기반의 방법으로는 총 1,965개의 추출된 키워드 중에는 943

5) 두 명의 독일어 전문가중 한 명은 논문의 저자이며, 다른 한 명은 독일어 학습 경력 10년 이상의 대학원 박사과정 학생이었음. 그러나 두 명 모두 자동차 기술분야의 전문가는 아니므로 이들에 의해 동시에 채택된 단어는 전문용어가 아닌 전문용어후보로 간주한다.

개의 전문용어후보가 포함되어 있어 약 48%의 정확률을 기록하였다. 단어리스트 기반 방법론으로 추출한 1,945개의 키워드 중에는 총 1,018개의 전문용어후보가 포함되어 있어 약 52.4%의 정확률을 기록하였다(표3).

최대 약 52%의 정확률은 타연구결과들과 비교하여 그리 높은 편은 아니라고 할 수 있지만 다른 연구결과들은 복잡한 절차와 프로그래밍들을 필요로 한다는 점에서 본 연구의 의의가 있다고 본다. 예를 들어 Hong et al. (2001)의 연구에서는 독일어 기술문서로부터 전문용어후보를 추출할 때 75%의 정확률을 얻을 수 있다고 보고하였다. 그러나 이 방법을 적용하기 위해서는 원문의 형태소 태깅을 위한 독일어 형태소 태거가 필요하고, n 그램 추출 및 통계 처리를 위한 프로그래밍의 절차가 필요하다.

<표 3> 전문용어후보 추출 정확률

	레퍼런스코퍼스기반	단어리스트기반
전문용어후보갯수	943	1,018
정확률	48%	52.4%

이와는 달리 본 방법론은 프리웨어인 'AntConc'와 역시 자유롭게 사용할 수 있는 단어리스트만 있으면 언제 어디서나 손쉽게 문서로부터 전문용어후보를 추출할 수 있다는 장점이 있다. 또한 어떠한 레퍼런스 코퍼스 또는 단어리스트를 사용하느냐에 따라 추출 정확률이 더 향상될 가능성이 있으므로 통번역가들에 의해 유용하게 사용될 수 있다.

레퍼런스 코퍼스 기반과 단어리스트 기반 방법론의 추출결과 중 가장 로그우도비가 큰 10개의 전문용어후보는 각각 다음 표4와 같았다.

<표 4> 상위 10개 전문용어후보

순위	레퍼런스코퍼스기반	단어리스트기반
1	pumpe	pumpe
2	duse	duse
3	kurbelwelle	kurbelwelle
4	ventil	ventil

5	passat	motorsteuergerät
6	motorsteuergerät	steuergerät
7	steuergerät	kraftstoffpumpe
8	kraftstoffpumpe	kolben
9	dichtung	abgasrückführung
10	schrauben	abstutzelement

두 방법 모두 상위 단어들의 차이는 크게 없었지만 정확률의 측면에서는 단어리스트 기반 방법론이 약 4.4% 정도 앞섰다. 그러나 이 결과를 가지고 두 방법론 중 어떠한 방법론이 더 우위에 있다고 주장하기는 어려울 것 같다. 레퍼런스 코퍼스 기반에서는 어떠한 코퍼스를 사용하느냐, 단어리스트 기반 방법론에서는 어떠한 단어리스트를 사용하느냐가 각 방법론의 성능을 결정짓는 중요한 요인이 될 것이기 때문이다. 본 연구에서는 비교적 소규모의 레퍼런스 코퍼스만을 사용하였는데 좀 더 규모가 크고 대표성이 있는 코퍼스를 사용한다면 추출 정확률은 달라질 것으로 예상된다. 마찬가지로 단어리스트 기반의 방법론에서도 어떠한 단어리스트를 사용하느냐에 따라 정확률은 달라질 것이다.

본 연구에서 주장하는 것은 어떠한 방법론이 우위에 있느냐가 아니라 손쉽게 접할 수 있는 프로그램과 언어리소스가 통번역가들이 매우 유용하게 사용할 수 있는 전문용어후보추출 도구의 역할을 대신할 수 있다는 점이다. 그 실험 결과가 비록 첨단 연구결과와 같이 75% 이상의 정확률을 보이는 것은 아니지만, 최소 50% 정도의 정확률이라 할지라도 실제 통번역을 준비하는 상황에서는 상당히 유용하게 사용될 수 있다는 것이다. 이 방법론을 통해 자동차분야 기술문서로부터 자동으로 추출된 전문용어후보 리스트는 부록에서 일부 소개한다.

Ⅵ. 맺는말

전문분야 통번역의 품질을 결정짓는 매우 중요한 요소 중의 하나는 전문용어의 자연스러운 통번역이다. 이를 위해 통번역 프로젝트를 착수하기 이전에 전문용어의 데이터베이스를 구축해놓고 작업을 시작한다. 전문용어 데이터베이스 구

축을 위해서는 해당 분야의 텍스트로부터 전문용어를 추출하여 이에 대한 역어를 준비해놓아야 하는데 방대한 분량의 텍스트로부터 전문용어를 추출한다는 것은 매우 큰 시간과 비용을 요하는 작업이다.

이를 위해 이미 상용 소프트웨어가 개발되어 있으나 상당한 비용과 사용방법에 대한 부담으로 독일어 관련 통번역 분야에서는 널리 사용되고 있지는 못한 실정이다. 이러한 문제를 해결하고자 본 연구에서는 통계언어학 연구를 위해 널리 사용되고 있는 'AntConc' 프로그램의 키워드 추출 기능을 사용해 전문용어후보를 손쉽게 추출할 수 있음을 보였다. 키워드를 추출하기 위해 필요한 레퍼런스 코퍼스 및 레퍼런스 단어리스트가 추출 정확률에 미치는 영향도 관찰하였다.

본 연구에서 제안한 방법론에 대한 검증작업에서는 자동차분야의 기술문서로부터 자동으로 추출한 전문용어후보가 최대 52.4%의 정확률을 보였다. 이는 최신 연구결과와 비교해서는 비교적 낮은 수치이지만 형태소 태거 등을 전혀 사용하지 않고 'AntConc'만을 사용해서 얻은 결과라는 점에서 실용적인 측면에서는 큰 의의가 있다고 본다.

이 연구가 더 큰 의미를 갖기 위해서는 추가적인 실험이 필요하다. 우선 레퍼런스 코퍼스의 규모와 성격에 따른 추출 정확률의 검토가 필요하다. 마찬가지로 단어 리스트의 규모와 성격에 따른 추출 정확률도 검토되어야 한다. 이를 위해 'IDS'의 'DeReWo'를 적용해보는 것도 필요해 보인다. 또한 본 연구에서는 전문용어후보를 한 단어 전문용어로만 한정하였는데 많은 경우 두 단어 이상으로 되거나 심지어는 구 Phrase의 형태로 되어 있는 전문용어가 존재하므로 이러한 전문용어후보의 추출을 위한 방법론도 간구되어야 할 것이다.

참고문헌

김정우(2003):「자연 과학 텍스트의 번역 방법론 시론」.『번역학연구』4권 1호, 27-49.

김성원, 김정우(2011):「전문용어 번역의 유형과 방법론: 의학 전문용어를 중심으로」.『번역학연구』12권 2호, 33-52.

방교영(2000):「러시아어 정보통신 전문용어 유형분석」.『통번역학연구』4권, 87-102.

홍문표(2008):「통제독일어가 번역수월성 향상에 미치는 영향에 대한 연구」.『독일언어문학』40권, 21-43.

Ahmad, K., A. Davies, H. Fulford & M. Rogers(1994): What is a term? The semi-automatic extraction of terms from text, Translation Studies: An Interdiscipline, 267-278.

Anthony, L.(2014): AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University.

Hansen-Schirra, S., S. Neumann & E. Steiner(2007): Cohesive explicitness and explicitation in an English-German translation corpus, Languages in Contrast 7.2, 241-266.

Heid, U.(1998): A linguistic bootstrapping approach to the extraction of term candidates from German text, Terminology 5.2, 161-181.

Heid, U.(1999): Extracting terminologically relevant collocations from German technical texts, TKE. Vol. 99, 241-255.

Hong, M., S. Fissaha & J. Haller(2001): Hybrid filtering for extraction of term candidates from German technical texts, TIA-2001, 223-232.

Jacquemin, C. & D. Bourigault(2000): Term Extraction and Automatic Indexing, Chapter 19 in Handbook of Computational Linguistics, Oxford University Press, Oxford.

Knorz, G.(1991): Indexieren, Klassieren, Extrahieren., in Grundlagen der praktischen Information und Dokumentation, München.

Kupietz, M., C. Belica, H. Keibel & A. Witt(2010): The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. in LREC, 1848-1854.

Pazienza, M., M. Pennacchiotti & F. Zanzotto(2005): Terminology extraction: an analysis of linguistic and statistical approaches, Knowledge Mining, 255-279.

Rayson, P. & R. Garside(2000): Comparing corpora using frequency profiling. Proceedings of the workshop on Comparing Corpora. Association for Computational Linguistics, 1-6.

Witschel, H.(2005): Terminology Extraction and Automatic Indexing, Terminology and Knowledge Engineering, 1-13.

Zusammenfassung

Extraktion der Termkandidaten aus deutschen technischen Texten mithilfe von 'AntConc'

Hong, Mun-Pyo (Sungkyunkwan Uni)

Sogar für die erfahrensten Übersetzer stellen technische Terme oft die größten Schwierigkeiten bei der Übersetzung technischer Dokumente dar. Die Schwierigkeiten der Übersetzung beruhen zum Teil darauf, dass es keinen passenden Begriff in der Zielsprache gibt, so dass es schwierig ist, diesen Begriff in der Zielsprache zu verbalisieren. Sie beruhen zum Teil auch darauf, dass dem Übersetzer oft das Fachwissen für die Übersetzung fehlt.

Aus diesem Grunde wird dem Übersetzer oft empfohlen, die Termdatenbank für eine Domäne im Voraus zu erstellen, um die korrekte und konsistente Übersetzung der Terme zu gewährleisten. Um eine Termdatenbank zu bilden, muss man aber zuerst die Terme aus technischen Texten extrahieren. Wenn der Umfang der zu übersetzenden Texte zu groß ist, dann ist es oft sehr mühsam und teuer, alle Terme zu entdecken. Viele Übersetzer sind oft für die Termextraktion und die Termdatenbankerstellung auf kommerzielle Produkte angewiesen.

Die vorliegende Arbeit stellt eine neue Methode vor, die für die Extraktion der Termkandidaten aus technischen Dokumenten angewendet werden kann. Diese Methode verwendet eine Freeware namens „AntConc“, die sich im Internet einfach und kostenlos herunterladen lässt. Eine Funktion der Software ist die Extraktion der Schlüsselwörter aus einem Text. Der Algorithmus der Funktion stützt sich auf das Log-likelihood Ratio Rechnen. Die zugrundeliegende Idee dieser Methode ist, dass ein Wort ein sehr wahrscheinlicher Termkandidat ist, wenn es relativ oft in einem Fachtext vorkommt, während es in einem allgemeinen Korpus aber relativ wenig vorkommt. Die Relativfrequenz wird hier durch das Log-likelihood Ratio Rechnen kalkuliert.

Das Experiment zeigt, dass der vorgeschlagene Ansatz etwa 52.4% Korrektheit aufweist. Den kommerziellen Systemen, die die sogenannte cutting-edge Technologie heranziehen, unterliegt unser Ansatz in der Korrektheit der Extraktion. Aber dieser Ansatz benötigt keine andere Ressourcen oder Tools als AntConc und ist vor allem kostenlos verwendbar, so dass er von jedem Übersetzer einfach benutzt werden kann.

핵심어 : 전문용어후보 Termkandidate, 상대빈도 Relativfrequenz,
AntConc AntConc, 키워드 Schlüsselwort,
로그우도비 Log-likelihood ratio

필자 E-mail : skkhmp@skku.edu

논문투고일 : 2015. 4. 15 / 심사일 : 2015. 5. 30 / 게재확정일 : 2015. 6. 5

〈부록: 자동추출 전문용어 후보 샘플 100개〉

레퍼런스 코퍼스 / 단어 리스트 공통 추출 단어	
abbremsen	federbein
abdichtung	federbeinachse
abflußpropfen	federbeinen
abgaskanal	federscheiben
abgaskrummer	federtellern
abgasrelevanten	federweg
abgasruckfuhrung	fehlerauslesegerat
abgasruckfuhrungsventil	fehlerauslesegerates
aggregateeinbau	fehlerspeicher
aggregatetrager	gummibuchsen
aggregatetragers	gummidichtung
airbagschlusselschalter	gummilager
airbagsteuergerat	gummilagerungen
aktivkohlebehalter	gummimetalllager
aktivkohlebehalterentluftung	gummitulle
anlassdrehzahl	handschaltgetriebes
anlaßdrehzahl	handschaltniveau
anlasserdrehzahl	handschuhfach
anlaufhalbringe	handschuhfaches
ansaugen	hauplagerdeckel
ansaugleistung	hauptbremszylinder
ansaugluft	hauptlager
ansaugluftmenge	hauptlagerdeckel
ansaugluftstrom	schaltgetriebe
ansaugrohr	schaltseilzuge
antriebsmotor	schalttafel
antriebsnockenspitze	schaumteil
antriebsrad	scheibenwischerarme
antriebsscheibe	schwinghebel
antriebsseil	schwinghebelachse
antriebswelle	schwingungen
antriebswellen	schwungrad
antriebswellengelenke	schwungradschrauben
anzeigefeld	selbststudienprogramm

drehpunkt	selbsttragende
drehrate	serienausstattung
drehrichtung	servolenkung
drehsteller	sicherungshalter
drehwinkel	sicherungsring
drehzahl	sideguard
drehzahlbereich	thermostat
drehzahlen	thompson
dreizylinder	tieflauf
drosselklappe	touran
drosselklappensektor	trennfuge
drosselklappenstutzen	triebwerk
einspritzanlage	ventilbetatigung
einspritzdrucken	ventilhub
einspritzmenge	ventilkopf
einspritzpumpe	zylinderkopfdeckel